FACULTAD DE INGENIERÍA MECÁNICA ELÉCTRICA, ELECTRÓNICA Y SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

# OPTIMIZACIÓN DEL PROCESO DE GESTIÓN DE PORTAFOLIO CREDITICIO CON LA IMPLEMENTACIÓN DE UN SISTEMA DE GESTIÓN UTILIZANDO DATA MINING.

## TESIS

PRESENTADO POR:

**MARCO ANTONIO CHUCUYA TIJUTANI**

**CRISTHIAN PERCY ALVARO MAMANI**

PARA OPTAR EL TÍTULO PROFESIONAL DE:

**INGENIERO DE SISTEMAS**

PUNO – PERÚ

2015

# UNIVERSIDAD NACIONAL DEL ALTIPLANO

FACULTAD DE INGENIERÍA MECÁNICA ELÉCTRICA, ELECTRÓNICA Y SISTEMAS

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

"OPTIMIZACIÓN DEL PROCESO DE GESTIÓN DE PORTAFOLIO CREDITICIO CON LA IMPLEMENTACIÓN DE UN SISTEMA DE GESTIÓN UTILIZANDO DATA MINING"

Tesis presentado por los bachilleres: MARCO ANTONIO CHUCUYA TIJUTANI
: CRISTHIAN PERCY ALVARO MAMANI

Para optar el título Profesional de : INGENIERO DE SISTEMAS

APROBADO POR

Presidente : Dr. Angel Manuel Olazabal Guerra

Primer Miembro : M.Sc. Elmer Coyla Idme

Segundo Miembro : M.Sc. Juan Antonio Flores Moroco

Director : M.Sc. Robert Antonio Romero Flores

Asesor : Ing. Adolfo Carlos Jiménez Chura

ÁREA: Informática
TEMA: Sistemas de información tradicionales y expertos

## AGRADECIMIENTO

A nuestros padres por sus ganas de vivir, sus consejos, cariño y comprensión. Todo lo que soy y lo que hasta en este momento he logrado es gracias Ustedes, a su sacrificio y esfuerzo, gracias por haberme impulsado en este trayecto de mi vida

A todas las personas del trabajo que estando cerca de nosotros y según su posibilidad nos tendieron una mano.

A los docentes de la Universidad Nacional del Altiplano, de la carrera Profesional de Ingeniería de Sistemas; por habernos inculcado de muchas capacidades bajo sus enseñanzas para sobresalir en esta etapa profesional.

En Memoria de Candelaria Tijutani Alanguia.

# DEDICATORIA

Marco, a mis padres Antonio Seúl Chucuya Zaga, y Candelaria Tijutani Alanguia, por todo el sacrificio, motivación y consejos que me brindaron en esta etapa importante de mi vida profesional

Cristhian, a mis padres Percy Alvaro Medina y Ana Luz Mamani Anccori , por todo el sacrificio, motivación y consejos que me brindaron en esta etapa importante de mi vida profesional.

Al Ing. Robert Antonio Romero Flores por la motivación dentro de temas importantes dentro de nuestra formación profesional.

A ustedes les dedicamos este trabajo y logro que representa una parte inicial de una vida profesional deseable para cualquier persona llena de ambiciones

# ÍNDICE

**CAPITULO I** 17

**PLANTEAMIENTO DEL PROBLEMA** 17

1.1 DEFINICIÓN DEL PROBLEMA 18
1.2 FORMULACIÓN DEL PROBLEMA 20
1.3 JUSTIFICACIÓN DE LA INVESTIGACIÓN 20
1.4 OBJETIVOS 22
1.5 HIPÓTESIS DE LA INVESTIGACIÓN 22
1.6 LIMITACIÓN DE LA INVESTIGACIÓN 23

**CAPITULO II** 24

**MARCO TEÓRICO** 24

2.1 ANTECEDENTES DE LA INVESTIGACIÓN 25
2.2 SUSTENTO TEÓRICO 27
2.3 DEFINICIÓN DE TÉRMINOS BÁSICOS 79
2.4 OPERACIONALIZACIÓN DE VARIABLES 80

**CAPITULO III** 82

**DISEÑO METODOLÓGICO DE LA INVESTIGACIÓN** 82

3.1 DISEÑO DE LA INVESTIGACIÓN 83
3.2 POBLACIÓN Y MUESTRA 83
3.3 TÉCNICAS E INSTRUMENTOS PARA RECOLECTAR LOS DATOS 84
3.4 MÉTODOS DE TRATAMIENTO DE DATOS 86
3.4.1 MATERIAL EXPERIMENTAL 86

**CAPITULO IV** 88

**ANÁLISIS E INTERPRETACIÓN DE RESULTADOS DE LA INVESTIGACIÓN** 88

4.1 CASO DE ESTUDIO 89
4.1.1 MODELO DE DATOS DE LA EMPRESA 89
4.1.2 NECESIDADES DE TECNOLOGÍAS DE INFORMACIÓN 90
4.1.3 RECURSOS TECNOLÓGICOS CON LOS QUE CUENTA 90
4.1.4 RECURSOS TECNOLÓGICOS QUE NECESITA 91
4.1.5 ARQUITECTURA DEL SISTEMA 91

4.1.6 MODELO DE DATOS MULTIDIMENSIONAL 104
4.1.7 LIMPIEZA E INTEGRACIÓN DE DATOS 108
4.1.8 REPORTES NECESARIOS 110
4.1.9 DISCUSIÓN FINAL 110
4.2 IMPLEMENTACIÓN 111
4.2.1 CREACIÓN DE CAPA DE INTEGRACIÓN 111
4.2.2 CREACIÓN DE CAPA DE ANÁLISIS 114
4.3 OPTIMIZACIÓN DEL PRE PROCESAMIENTO DE DATOS. 122

**CONCLUSIONES 126**

**RECOMENDACIONES 127**

**BIBLIOGRAFÍA 128**
**ANEXOS 130**

# ÍNDICE DE TABLAS

**TABLA N° 1**: OFICINAS ESPECIALES DE CAJA LOS ANDES 29
**TABLA N° 2**: PRODUCTOS CREDITICIOS DE CAJA LOS ANDES 30
**TABLA N° 3:** APLICACIONES DE DATA MINING. 31
**TABLA N° 4:** GRADOS DE NORMALIZACIÓN 42
**TABLA N° 5**: CAMPOS DE LA BASE DE DATOS, INFORMACIÓN PROCESADA. 66
**TABLA N° 6:** PRESENTACIÓN DE FALLIDOS. 69
**TABLA N° 7:** FACTORES DE CALIDAD Y MÉTRICAS DE CALIDAD DE SOFTWARE. 74
**TABLA N° 8:** RESUMEN DE LAS CARACTERÍSTICAS GENERALES DEL SISTEMA (FUENTE: LEBRUN Y SANTILLAN, 2008) 75
**TABLA N° 9**: CUADRO DE OPERACIONALIZACIÓN DE VARIABLE 80
**TABLA N° 10:** DISEÑO DE INVESTIGACIÓN CUASIEXPERIMENTAL 83
**TABLA N° 11**: PERIODOS DE POBLACIÓN. 84
**TABLA N° 12**: PERIODOS DE MUESTRA. 84
**TABLA N° 13**: PROCESOS PARA LA ELABORACIÓN DE REPORTES DE COSECHAS, COMO PARTE A OPTIMIZAR DE LA GESTIÓN DE PORTAFOLIO CREDITICIO. 85
**TABLA N° 14:** ARQUITECTURA EN CADA FRAMEWORK 93
**TABLA N° 15:** SERVICIOS IDENTIFICADOS PARA EL SISTEMA PROPUESTO 100
**TABLA N° 16:** VALORES DE LOS ATRIBUTOS DEL MODELO MULTIDIMESIONAL 101
**TABLA N° 17:** OCLIA01 - CLIENTES - SIFCNET 104
**TABLA N° 18:** SCONA05 - OFICINAS - SIFCNET 107
**TABLA N° 19**: OCLIA19 - ASESORES – SIFCNET 107
**TABLA N° 20:** OCREA02 - CATEGORÍA PRODUCTOS ACTIVOS - SIFCNET 108
**TABLA N° 21**: OCREA03 - PRODUCTOS ACTIVOS – SIFCNET 108
**TABLA N° 22**: ESTÁNDAR DE NOMBRES DEFINIDO PARA LA IMPLEMENTACIÓN DE DATAWAREHOUSE 109

## ÍNDICE DE GRÁFICOS

**GRAFICO N° 1**: COMPORTAMIENTO CANÓNICO DE FALLIDOS 70
**GRAFICO N° 2:** COMPORTAMIENTO ESTABLE DE FALLIDOS. 71
**GRAFICO N° 3:** COMPORTAMIENTO NO ESTABLE DE FALLIDOS. 71
**GRAFICO N° 4:** RESULTADOS EN LA PRUEBAS CON SGP, EN EL PROCESO ETL INMERSO. 125

## ÍNDICE DE FIGURAS

**FIGURA N° 1:** COMPONENTES DE LA GESTIÓN DE PORTAFOLIO (ENFOCADO COMO PROCESO). 18
**FIGURA N° 2:** REPORTE DE PORTAFOLIO EN HOJAS DE CÁLCULO. 19
**FIGURA N° 3:** DISCIPLINAS RELACIONADAS A DATA MINING. 32
**FIGURA N° 4:** DW Y DATA MARTS. 34
**FIGURA N° 5:** ARQUITECTURA DEL PROCESO DATA WAREHOUSING. 39
**FIGURA N° 6**: DIAGRAMA DE CASO DE USO. 45
**FIGURA N° 7:** DIAGRAMA DE SECUENCIA. 46
**FIGURA N° 8:** DIAGRAMA DE COLABORACIÓN. 47
**FIGURA N° 9:** DIAGRAMA DE ESTADOS. 49
**FIGURA N° 10**: ESQUEMA DE 4 NIVELES DE CRISP-DM 50
**FIGURA N° 11**: MODELO DE PROCESO CRISP-DM. 52
**FIGURA N° 12**: FASE DE COMPRENSIÓN DEL NEGOCIO. 53
**FIGURA N° 13**: FASE DE COMPRESIÓN DE LOS DATOS. 55
**FIGURA N° 14**: FASE DE PREPARACIÓN DE LOS DATOS 57
**FIGURA N° 15**: FASE DE MODELADO 59
**FIGURA N° 16**: FASE DE EVALUACIÓN. 63
**FIGURA N° 17**: FASE DE IMPLEMENTACIÓN. 64
**FIGURA N° 18:** DINÁMICA DE DATOS POR EL TRANSCURSO DE TIEMPO. 68
**FIGURA N° 19:** AGRUPACIÓN A SALDO VENCIDO. 68
**FIGURA N° 20:** AGRUPACIÓN A MONTO DESEMBOLSADO 68
**FIGURA N° 21:** FACTORES QUE AFECTAN LA CALIDAD DE SOFTWARE. (FUENTE: PRESSMAN, 2002). 73
**FIGURA N° 22**: FLUJO DEL PROCESO PRE SGP DE GESTIÓN DE PORTAFOLIO CREDITICIO – GENERACIÓN DE COSECHAS CREDITICIAS 85
**FIGURA N° 23**: FRAMEWORK ARQUITECTÓNICO 93
**FIGURA N° 24:** MODELO DE DOMINIO 94
**FIGURA N° 25**: ACTORES DEL SISTEMA. 94
**FIGURA N° 26**: MODELO DE CASOS DE USO DEL SISTEMA 96
**FIGURA N° 27:** DIAGRAMA DE CASO DE USO, CARGA DE DATOS SIFCNET 96
**FIGURA N° 28:** DIAGRAMA DE CASOS DE USO, EJECUCIÓN DE PROCESO ETL 97
**FIGURA N° 29**: DIAGRAMA DE CASOS DE USO, CONSULTA EN EL SGP 97
**FIGURA N° 30:** DIAGRAMA DE CASOS DE USO, CONSULTAS AL SGP NIVEL ADMINISTRADOR 98
**FIGURA N° 31:** PATRÓN DE ARQUITECTURA PARA EL SISTEMA 99
**FIGURA N° 32:** ARQUITECTURA DE SGP. 100
**FIGURA N° 33**: MODELO DE DATOS 101

**FIGURA N° 34:** VISTA DE DEPLOYEMENT. 103
**FIGURA N° 35:** PROCESO DE CONSTRUCCIÓN DE ESTRUCTURA DW. 112
**FIGURA N° 36:** SCRIPTS SQL DE CONSTRUCCIÓN DE ESTRUCTURA DW. 112
**FIGURA N° 37:** PROCESO DE CARGA DE TABLA DE HECHOS. 113
**FIGURA N° 38**: PROCESO DE CARGA DE TABLAS DIMENSIONES. 113
**FIGURA N° 39:** PROGRAMACIÓN DEL PROCESO DE CARGA DE DWH. 114
**FIGURA N° 40:** SCHEMA WORKBENCH CUBO OLAP COSECHAS. 115
**FIGURA N° 41:** SCHEMA WORKBENCH XML QUE REPRESENTA AL CUBO OLAP. 116
**FIGURA N° 42:** SCHEMA WORKBENCH PUBLICACIÓN DE CUBO OLAP A PENTAHO **SERVER-BI.** 116
**FIGURA N° 43:** INICIO DE SESIÓN EN PENTAHO SERVER-BI. 118
**FIGURA N° 44:** ANÁLISIS DE DATOS CON SAIKU ANALYTICS**.** 118
**FIGURA N° 45:** ANÁLISIS DE DATOS CON SAIKU ANALYTICS**.** 119
**FIGURA N° 46:** ANÁLISIS DATOS CON SAIKU MODELAMIENTO DE LOS DATOS**.** 119
**FIGURA N° 47**: ANÁLISIS DATOS CON SAIKU ANALYTICS GENERACIÓN DE CONOCIMIENTO. 120
**FIGURA N° 48**: ANÁLISIS DATOS CON SAIKU ANALYTICS. 120
**FIGURA N° 49**: ANÁLISIS DATOS CON SAIKU ANALYTICS. 121
**FIGURA N° 50**: ANÁLISIS DATOS CON SAIKU GENERACIÓN DE CONOCIMIENTO Y MODELAMIENTO DE REPORTES. 121
**FIGURA N° 51**: ANÁLISIS DATOS CON SAIKU ANALYTICS. 122
**FIGURA N° 52**: ANÁLISIS DATOS CON SAIKU ANALYTICS. 122

## ÍNDICE DE ANEXOS

**ANEXO 1**: ENCUESTA DE PRE – PRUEBA. 131
**ANEXO 2:** MATRIZ DE CONSISTENCIA 132

**RESUMEN**

El presente trabajo de investigación denominado "OPTIMIZACIÓN DEL PROCESO DE GESTIÓN DE PORTAFOLIO CREDITICIO CON LA IMPLEMENTACIÓN DE UN SISTEMA DE GESTIÓN UTILIZANDO DATA MINING", se desarrolló con el propósito de implementar una solución de inteligencia de negocios en el proceso operativo de gestión de portafolio que se desarrolla en la CAJA RURAL DE AHORRO Y CRÉDITOS, LOS ANDES, específicamente en la Gerencia de Riesgos. Verificando un estado actual y evaluando los cambios que pueden ocasionar implementar el Sistema de Gestión de Portafolio Crediticio utilizando Data Mining, dando como principales actores a los analistas de riesgos, Gerentes, Jefes y encargados, con encuestas, tomando un histórico del proceso anterior utilizado y verificando la eficacia que posee comparándolo con la nueva propuesta. Además de emplear técnica de métricas de medición de eficacia para evaluar la mejora de respuesta y eficacia, en los procesos. El objetivo general de esta tesis es la optimización del proceso, en un tiempo determinado y la robustez en respuesta. Utilizando un diseño e implementación de un prototipo que integre las tecnologías de inteligencia empresarial: Datawarehousing, OLAP y data Mining, los cuales ofrecen un ambiente integral y factible para la necesidades de un corporativo con datos masivos.

**PALABRAS CLAVE:** Portafolio Crediticio, Micro Finanzas, Inteligencia de Negocios, Riesgos, Riesgo de Crédito.

**ABSTRACT**

The present research work called " OPTIMIZATION PROCESS MANAGEMENT PORTFOLIO CREDIT TO THE IMPLEMENTATION OF MANAGEMENT SYSTEM USING DATA MINING " , was developed with the purpose of implementing a business intelligence solution in the operating portfolio management process that develops in the RURAL SAVINGS AND LOANS , LOS ANDES , specifically in Risk Management . Verifying a current state and evaluating changes that can cause implement Management System Credit Portfolio using Data Mining , leading to key players at risk analysts , managers, bosses and managers, with surveys, taking a record of the previous process used and verifying the effectiveness it has compared to the new proposal. Addition to using metric measuring technique to evaluate the effectiveness of response and efficiency improvement in the processes . The overall objective of this thesis is the optimization of the process, in a certain time and robustness in response . Using a design and implementation of a prototype that integrates business intelligence technologies : Data warehousing , OLAP and data Mining , which offer a comprehensive and feasible for the needs of a corporate environment with massive data .

**KEY WORDS:** Credit Portfolio, Micro Finance, Business Intelligence, Risk Management, Credit Risk.

## INTRODUCCIÓN

La Caja Rural de Ahorro y Crédito Los Andes, es una Institución del Sistema Financiero regulada por Superintendencia de Banca, Seguro y AFP's; que impulsa el desarrollo en los micros empresarios o medianos empresarios rurales. En un contexto de gestión de riesgos, la empresa rural no califica para las Financieras de nivel superior por su capital social, tales como los bancos, sin embargo el modelo de tecnología crediticia que viene aplicando la Caja Rural de Ahorro y Credito Los Andes, permite acceder a un crédito a mencionados empresarios. Es así que la Institución en 19 años de funcionamiento, desempeñando el papel de impulsor y fondeador a varias empresas rurales. Llego a emitir más de 150 millones en créditos, en más de 35 mil clientes de las Regiones de Puno, Ayacucho, Cuzco, y Arequipa. En las que cuenta con Oficinas de Atención directa al público.

En un mundo corporativo competitivo donde prima las buenas prácticas de estándares y certificaciones, en el cual esta relevantemente considerado la Gestión de Riesgos y dentro de este la Inteligencia de Negocios como herramienta que tiene la función principal de dar significado a los enormes repositorios de datos que se genera en la empresa. No *explotar* la información significa la disminución de la agresividad frente a la competencia.

El crecimiento de la Institución radica en las colocaciones y la bancarización de más empresarios rurales. El papel que desempeña la Inteligencia de Negocios al dar tratamiento y muestreo a los datos en variados modos, es indicar la situación a nivel panorámico de todos los créditos (en este contexto) de manera

que ayude a la toma de decisiones de los principales ejecutivos de la Institución para alinear el propósito y objetivo que se tiene planteado como corporación.

Los indicadores de Gestión de Riesgos, son ratios que permiten percibir a priori la situación actual de la Institución tanto positivos como negativos, de esta manera establecer la tolerancia al riesgo que se mantiene. El indicador de Fallidos y Tasa de Mora Prematura (TMP) son parte de indicares clave de la gestión de riesgos, que indican cuan temprano un crédito tiende a presentar mora y cuanto impacta esto a los resultados financieros, además de otras implicancias como controles en los procesos que llevaron a dicha situación.

El mantener el control y tratamiento en cada caso de comportamiento de los Indicadores permite establecer una adecuada administración de los riesgos, optando por mitigarlos y así alcanzar los objetivos de crecimiento y social planteados por la Institución.

En el presente proyecto de investigación se ha definido cuatro capítulos, los cuales se estructuran de la siguiente manera:

**Capítulo I, Planteamiento del Problema:** Describe el problema objeto de estudio, los objetos planteados para la resolución del mismo.

**Capitulo II, Marco Teórico:** Contiene los antecedentes relacionados al tema de investigación, las bases teóricas necesarias para el desarrollo del estudio.

**Capitulo III, Diseño Metodológico de la Investigación:** Refleja la metodología empleada para el desarrollo del estudio, tipo de investigación, población y muestra, técnicas e instrumentos de recolección de datos aplicados.

**Capitulo IV, Análisis e Interpretación de Resultados de la Investigación:** Trata sobre los procesos que permiten analizar la información recopilada; verificar su confiabilidad mediante la triangulación; interpretar y comprender los resultados; y presentar y usar los resultados.

# CAPITULO I

# PLANTEAMIENTO DEL PROBLEMA

## 1.1 DEFINICIÓN DEL PROBLEMA

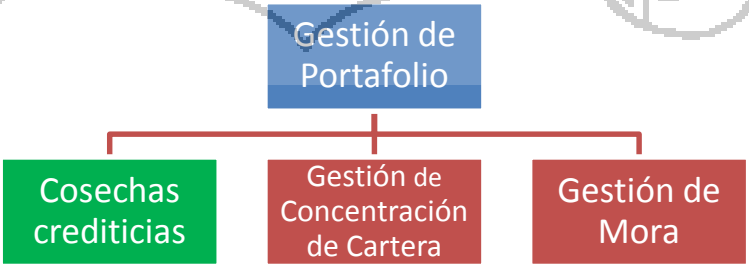
La Caja de Ahorro y Crédito Los Andes, Institución Financiera pionera en créditos rurales en la Región de Puno, viene colocando créditos de manera que estos sirvan de apoyo para el desarrollo del emprendedor rural, ofreciendo sus diferentes productos evaluados acorde a la región. Enfrentando riesgos tales como Crediticio, Liquidez y Mercado, Operacional; como los más comunes.

La Oficina de Riesgos tiene como misión mitigar estos riesgos que a posterior cada riesgo deriva a generar pérdidas financieras, denotando que el riesgo de más perjuicio es el Crediticio que infiere directamente en las perdidas al significar el incumplimiento de pago por parte del deudor.

En los últimos años se ha observado que el 23% de perdidas debido al incumplimiento de las responsabilidades del deudor, significándose esto un riesgo latente el cual debe ser mitigado.

La SBS en la Resolución 3780-2008 plantea que cada Institución Financiera implementara un Sistema de "Cosechas", que es parte del Sistema de Portafolio crediticio, y también se constituye como parte del proceso de Gestión de Portafolio.

***Figura N° 1*:** Componentes de la Gestión de Portafolio (enfocado como proceso).



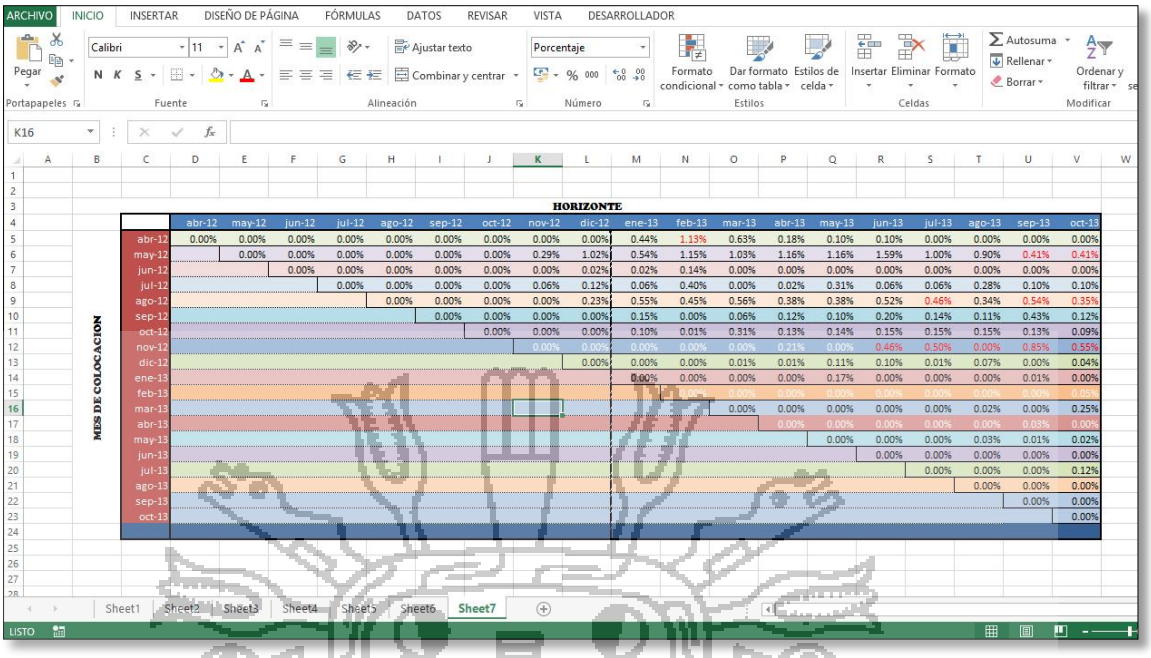

**Elaboración:** Por los investigadores

La gestión de portafolio crediticio con su sub proceso Cosechas crediticias; se constituye como una herramienta para mitigar el riesgo de crédito, debido a que su función permite la verificación de series de créditos agrupados por un elemento en particular, el más notable es por periodo. Entonces cada crédito como tal posee independientemente un comportamiento de saldo o lo que resta por pagar, en caso tenga demoras en cierto periodo este mostrara un indicador como señal para tomar acciones de empresa para evitar pérdidas.

Bajo la referencia del párrafo anterior, un portafolio de créditos tal como se conceptúa; contiene un número mayoritario de créditos o registros de créditos, y estos segmentados por el elemento tiempo, tienden a multiplicarse, esto debido a que existen datos dentro del registro de crédito que varían con el tiempo, y a posterior esto causa el incremento de registros.

Actualmente la Caja Rural de Ahorro de Crédito Los Andes, no tiene implementado un Sistema de Portafolio de Créditos adecuado; se viene contemplando este requerimiento del ente supervisor, con hojas de cálculo de uso común, que no soporta una gestión de datos eficaz además de que sus uso es rudimentario y manual, sujeto a errores humanos.

***Figura N° 2:*** *Reporte de Portafolio en hojas de cálculo.*




**Elaboración:** Por los investigadores.

Con el presente trabajo de investigación se pretende desarrollar un sistema dedicado exclusivamente para la gestión de datos que requiere la gestión de portafolio de créditos y sus respectivos reportes para informes del analista, es por ello que se plantea la formulación del problema tal como sigue.

## 1.2 FORMULACIÓN DEL PROBLEMA

¿Qué efectos produce el Sistema de Gestión de Portafolio Crediticio con el uso de Data Mining en la gestión de portafolio de la Caja Rural de Ahorro y Crédito Los Andes?

## 1.3 JUSTIFICACIÓN DE LA INVESTIGACIÓN

Actualmente el sector crediticio en el país se ha convertido en un impulsor principal del crecimiento económico, y debido a su accesibilidad dada por la oferta financiera, es el socio inmediato de un micro, pequeño o mediano empresario. Por lo que internamente una empresa del sistema financiero tendrá

que gestión los datos de cada operación crediticia, en tantos más créditos se otorgue más datos tendrá que albergar y procesar.

A la vez consideremos a todos los datos gestionados como información, muy aparte de mantener un orden y gestión de créditos, la información representa una herramienta clave para poder sobrevivir en un mercado cambiante, dinámico y global. Aprender a competir con esta información es fundamental para la toma de decisiones, el crecimiento y la gestión de nuestra empresa. La disciplina denominada como Business Intelligence nos acerca a los sistemas de información que nos ayudan a la toma de decisiones en nuestra organización. La pyme dispone, como todas las empresas, no importa su tamaño, de sistemas de información más o menos sofisticados y que es conveniente analizar y optimizar.

Dentro de los sistemas de información que presenta la Business Intelligence, se contempla aquellos que ayuden a la toma de decisiones con la minería de datos o data mining.

El presente trabajo de investigación, a través de un sistema que utilice la tecnología de data mining, permitirá proponer una alternativa para la gestión de portafolio de créditos de la Caja Rural de Ahorro y Crédito Los Andes; que mediante la utilización de dicha aplicación, los analistas financieros, de riesgos y gerentes se beneficiarán al contar con esta herramienta de toma de decisiones que le proporcionara reportes de comportamientos de series de créditos segmentados, luego de un proceso de miles de datos.

Además el trabajo de investigación permitirá conocer si la utilización de un sistema con tecnología de minería de datos; para la gestión de portafolio de

créditos, tiene un efecto positivo en el proceso de datos de los registros de créditos y la toma de decisiones ejecutivas.

## 1.4 OBJETIVOS

### 1.4.1 OBJETIVO GENERAL

Optimizar el proceso de gestión de portafolio crediticio con la implementación de un sistema de gestión utilizando la tecnología Data Mining para la elaboración de Cosechas Crediticias que derive en toma de decisiones en la Caja Rural de Ahorro y Crédito Los Andes.

### 1.4.2 OBJETIVOS ESPECÍFICOS

a) Determinar los requerimientos para el desarrollo del Sistema de Gestión de Portafolio Crediticio.

b) Identificar y optimizar el Pre procesamiento de datos para la Gestión de Portafolio Crediticio, en la elaboración de las Cosechas crediticias.

c) Determinar el Modelo de negocio acorde a los procesos de Gestión de Portafolio en la elaboración de las cosechas.

d) Generar Conocimiento como resultado del proceso de gestión aplicando Data Mining.

## 1.5 HIPÓTESIS DE LA INVESTIGACIÓN

### 1.5.1 HIPÓTESIS GENERAL

El sistema de gestión utilizando tecnología Data Mining optimiza produce efectos positivos en la gestión de portafolio crediticio para la toma de decisiones en la Caja Rural de Ahorro y Crédito Los Andes.

### 1.5.2 HIPÓTESIS ESPECÍFICAS

a) Se obtendrá los requerimientos de ejecutivos de la Caja Rural de Ahorro y Crédito Los Andes para el crecimiento corporativo, que viabilice el

desarrollo del sistema de gestión de portafolio crediticio en la elaboración de cosechas crediticias.

b) El proceso de preparación de datos realizados manualmente serán optimizados por el componente de Data Mining- ETL.

c) La metodología de desarrollo CRISP-DM permite el modelamiento de negocio para la implementación del Sistema de Gestión de Portafolio Crediticio.

d) El Sistema de Gestión de Portafolio Crediticio como resultado generara conocimiento para la toma de decisiones.

## 1.6 LIMITACIÓN DE LA INVESTIGACIÓN

El sistema comprende más gestión de datos que manejo y jerarquías de interfaces. Tiene la opción de visualización de reportes bajo filtros o parámetros; como entregable final.

Solo determinadas áreas de la empresa podrán ser los usuarios del sistema. Se necesita de formación especializada para el entendimiento de los entregables del sistema. Se aplicara íntegramente en la empresa Caja Los Andes, con sede corporativa ubicado en el distrito de Puno, región Puno.

# CAPITULO II

# MARCO TEÓRICO

## 2.1 ANTECEDENTES DE LA INVESTIGACIÓN

Se presenta los antecedentes de investigaciones de tesis y proyectos relacionados con el tema de investigación.

### 2.1.1 PORTAL WEB DE NOTICIAS DEL PERÚ APLICANDO TÉCNICAS DE RECUPERACIÓN DE INFORMACIÓN Y MINERÍA DE DATOS

Realizado por Alain Alejo Huarachi. Dicha investigación tiene como objetivo general "Desarrollar un portal web de noticias del Perú aplicando técnicas de recuperación de información y minería de datos" y llegó a las siguientes conclusiones. "El portal web de noticias del Perú, que aplica técnicas de recuperación de información y minería de datos, logra extraer, clasificar y visualizar las noticias publicadas en otros sitios web de diversos medios de comunicación del Perú".

### 2.1.2 UTILIZACIÓN DE LA TECNOLOGÍA DATA WAREHOUSE EN INSTITUCIONES EDUCATIVAS"

Realizado por Rene Cruz Guerrero. Dicha investigación tiene como objetivo general "Aplicar la Tecnología DataWarehouse en instituciones educativas, con la finalidad de que cuenten con un sistema que resuelva sus necesidades de tipo informacional en procesamiento OLAP en el ámbito de soporte para la toma de decisiones" y llegó a las siguientes conclusiones. "La selección de una herramienta que se utilice para el desarrollo de un DW no es fácil, debido a que se deben considerar diversos aspectos. Uno de ellos, es verificar que procesos permite realizar (extracción, transformación, carga y explotación de los datos). Algunas herramientas, permiten realizar un solo proceso, otras la combinación de algunos de ellos y otras todos los procesos.

Otros de los aspectos a considerar son: el tamaño del DW, la compatibilidad de plataformas y su costo".

### 2.1.3 DATA WAREHOUSING CON PROCESAMIENTO DE DATOS TEXTUALES

Realizado por Elizabet Tejeda Ávila. Dicha investigación tiene como objetivo general "La Definición e implementación de un nuevo modelo multidimensional, con soporte a tributos textuales en sus dimensiones" y llegó a las siguientes conclusiones. "Se ha formalizado matemáticamente el nuevo modelo multidimensional, que garantiza el procesamiento del conocimiento implícito de atributos textuales de base de datos obtenido de forma previa, en entorno OLAP".

### 2.1.4 INTELIGENCIA EMPRESARIAL PARA LA TOMA DE DECISIONES EN LA PYME ENFOCADA EN LA ADMINISTRACIÓN DE LA RELACIÓN CON EL CLIENTE (CRM) UTILIZANDO ANÁLISIS DE LA CANASTA DE COMPRA (MBA)

Realizado por Yesid Valentin Apaza Gonzales. Dicha investigación tiene como objetivo general "El incremento de la rentabilidad en un tiempo determinado utilizando un diseño e implementación de un prototipo que integre las tecnologías de inteligencia empresarial; datawarehousing, OLAP y data Mining, ofreciendo un ambiente integral y factible para las necesidades de una PyME comercializadora" y llegó a las siguientes conclusiones. "Se comprueba el incremento de rentabilidad en un 23.2% y la elaboración del prototipo con factibilidad de la aplicación del proceso de inteligencia empresarial, mediante la metodología propuesta, en las PyMES, lo que hasta ahora no se ha abordado con profundidad, por cuestiones de infraestructura y recursos limitados para El".

### 2.1.5 EXPERIENCIAS PRÁCTICAS EN LA MEDICIÓN DE RIESGO CREDITICIO DE MICRO EMPRESARIOS UTILIZANDO MODELOS DE CREDIT SCORING

Publicado por los estudiantes; Cristian Bravo, Sebastián Maldonado, Richard Weber.

Dicha publicación experimental, menciona que para los modelos de medición se tiene como fase el pre procesamiento de datos con el fin de adecuarlos al método de minería de data.

## 2.2 SUSTENTO TEÓRICO

### 2.2.1 Portafolio de Créditos.

O también denominado Portafolio Crediticio, es una colección de registros o tuplas de los datos de cada crédito pactado con la empresa financiera, donde contiene la información de los clientes o deudores contrayentes y los datos de control de saldos por cada crédito, entre otros datos.

#### 2.2.1.1 Crédito.

Implica una transacción de un valor presente, por una promesa de pago en un tiempo especificado en el futuro. En una transacción a crédito el comprador o deudor demuestra su poder o influencia para obtener el permiso del vendedor o acreedor para usar su capital. Consumada la transacción, se crea el derecho del vendedor a recibir el pago en el futuro y la obligación del comprador de pagar en el tiempo designado. La obligación de pagar es, a la vez moral y legal; las leyes de todo estado previenen la acción legal en contra del deudor moroso.

#### 2.2.1.2 Cartera Atrasada:

Es la suma de los créditos vencidos y en cobranza judicial.

#### 2.2.1.3 Cartera de Alto Riesgo:

Es la suma de los créditos reestructurados, refinanciados, vencidos y en cobranza judicial.

**2.2.1.4 Cartera Pesada:**

Es la suma de los créditos directos e indirectos con calificaciones crediticias del deudor de deficiente, dudoso y pérdida.

**2.2.1.5 Créditos Directos:**

Es la suma de los créditos vigentes, reestructurados, refinanciados, vencidos y en cobranza judicial. Los créditos en moneda nacional incluyen también los de valor de actualización constante. Para convertir los créditos en moneda extranjera se utiliza el tipo de cambio contable de fin de periodo.

**2.2.1.6 Créditos Castigados:**

Créditos clasificados como pérdida, íntegramente provisionados, que han sido retirados de los balances de las empresas. Para castigar un crédito, debe existir evidencia real de su irrecuperabilidad o debe ser por un monto que no justifique iniciar acción judicial o arbitral.

**2.2.1.7 Modalidad de Pago:**

Establece el intervalo de días dentro del plazo que debe ser pagado un crédito. Dependiendo de la evaluación de la capacidad pago, la rotación de inventarios y periodos de ingresos del deudor o contraparte.

**2.2.2 Producto de Crédito.**

Dependiendo al sector económico empresarial que va dirigido un crédito, este debe satisfacer características que logren cumplir con el requerimiento de mencionado sector. Este conjunto de características se engloban y tipifican al

crédito dirigido particularmente para el sector que se le asigno. Por tanto el paquete de características se establece como Productos de crédito.

### 2.2.3 Caja Rural de Ahorro y Crédito Los Andes S.A:

Empresa del Sistema Financiero, Regulado por la Súper Intendencia de Banca, Seguros y AFP's. Da los servicios financieros de Crédito y Ahorro, dirigido al ámbito rural. Creado el 19 de Noviembre de 1997, con funcionamiento en el Departamento de Puno, donde también reside la sede corporativa.

Actualmente posee más de 150 millones en créditos colocados, en 21 Oficinas Especiales (al corte del periodo Junio 2014). El expertise adquirido en el funcionamiento en el rubro, le ha impulsado a crear productos de crédito dirigidos al ámbito rural, esto los hace pioneros en la región, al implementar el producto Agropecuario.

***Tabla N° 1**: Oficinas Especiales de Caja Los Andes*

| Región | Oficina Especial |
|---|---|
| Puno | Puno |
| | Juliaca |
| | Ilave |
| | Desaguadero |
| | Ayaviri |
| | Azángaro |
| | Coata |
| | Huancané |
| | Acora |
| | Yunguyo |
| | Pedro Vilcapaza |
| | Macusani |
| | Taraco |
| Cusco | Espinar |

| | |
|---|---|
| | Sicuani |
| Arequipa | Aplao |
| | Chivay |
| Ayacucho | Huanta |
| | Pampa Cangallo |
| | Ayacucho |

**Fuente:** Caja Los Andes.

***Tabla N° 2**: Productos crediticios de Caja Los Andes*

| **Producto** |
|---|
| Agropecuario |
| Agrícola |
| PYME |
| Grupo Solidario |
| Contra deposito |
| Convenio |
| Libre Disponibilidad |

**Fuente:** Caja Los Andes.

### 2.2.3 Data Mining.

Es la exploración automática o semiautomática de grandes cantidades de datos para el descubrimiento de reglas y patrones.

Proceso iterativo de detección y extracción de patrones a partir de grandes bases de datos, modelo de reconocimiento.

Es el análisis de un conjunto de datos para encontrar relaciones desconocidas y resumir los datos de nuevas formas entendibles para el minero. Es el proceso analítico, por medio del cual se extrae información oculta de grandes cantidades de datos siendo muy útil para predecir futuros comportamientos y tendencias.

#### 2.2.2.1 Etapas del Data Mining.

a) **Determinación de los objetivos.** Trata de la delimitación de los objetivos que el cliente desea bajo la orientación del especialista en data mining.

b) **Pre procesamiento de los datos.** Se refiere a la selección, la limpieza, el enriquecimiento, la reducción y la transformación de las bases de datos. Esta etapa consume generalmente alrededor del setenta por ciento del tiempo total de un proyecto de data mining.

c) **Determinación del modelo.** Se comienza realizando unos análisis estadísticos de los datos, y después se lleva a cabo una visualización gráfica de los mismos para tener una primera aproximación. Según los objetivos planteados y la tarea que debe llevarse a cabo, pueden utilizarse algoritmos desarrollados en diferentes áreas de la Inteligencia Artificial.

d) **Análisis de los resultados.** Verifica si los resultados obtenidos son coherentes y los coteja con los obtenidos por los análisis estadísticos y de visualización gráfica. El cliente determina si son novedosos y si le aportan un nuevo conocimiento que le permita considerar sus decisiones.

### 2.2.2.2 Aplicaciones de Data Mining

***Tabla N° 3:*** *Aplicaciones de Data Mining.*

| Área de aplicación | Ejemplos |
|---|---|
| **Aplicaciones financieras** | Obtención de patrones de uso fraudulento de tarjetas de crédito, determinación del gasto en tarjeta de crédito por grupos, cálculo de correlaciones entre indicadores financieros, análisis de riesgo en créditos. |
| **Análisis de mercado, distribución y comercio** | Análisis de la canasta básica de mercado, evaluación de campañas publicitarias, análisis de la fidelidad de los clientes, estimación de inventarios, costos y ventas. |
| **Seguros y salud privada** | Determinación de clientes potencialmente caros, identificación de patrones de comportamiento para clientes con riesgo, identificación de comportamiento fraudulento, |

| | |
|---|---|
| | predicción de clientes que podrían ampliar su póliza. |
| **Educación** | Selección o captación de estudiantes, detección de abandonos y fracasos, estimación de tiempo de estancia en la institución. |
| **Procesos industriales** | Extracción de modelos sobre comportamiento de compuestos, detección de piezas con defectos, predicción de fallos y accidentes, estimación de composiciones óptimas en mezclas, extracción de modelos de costos, extracción de modelos de producción. |
| **Medicina, biología, bioingeniería y otras ciencias** | Diagnóstico de enfermedades, detección de pacientes con riesgo de sufrir una enfermedad concreta, recomendación priorizada de fármacos para una misma enfermedad, predecir si un compuesto químico causa cáncer, clasificación de cuerpos celestes, predicción del recorrido y distribución de inundaciones, modelos de calidad de aguas. |
| **Telecomunicaciones** | Establecimiento de patrones de llamadas, modelos de carga en redes, detección de fraude. |

***Figura N° 3:*** *Disciplinas relacionadas a Data Mining.*


**Fuente:** Presentación M.Sc. Robert Romero (2011).

### 2.2.4 Data Warehouse.

El concepto de Data Warehouse (almacén de datos), surge como una solución para obtener la información necesaria para la toma de decisiones, sin embargo, no es únicamente un almacén de datos, sino que su característica principal es la forma en como están estructurados esos datos, de modo que solucione cualquier tipo de consulta de manera eficiente u en el menor tiempo posible. A continuación, se muestran algunos conceptos de DW:

I. Un DW, es un repositorio de información coleccionada desde múltiples fuentes, bajo un esquema uniforme u que usualmente reside en un solo sitio (JIAWEI et. al, 2001).

II. Los DW son construidos vía n proceso de limpieza, transformación, integración u carga de datos.

Es una colección de datos orientados al sujeto, integrados, de tiempos variantes y no volátiles, que sirven de soporte para el proceso de toma de decisiones (MUKESH, 1999).

III. Es un almacenamiento de información homogénea u fiable, en una estructura basada en la consulta y el tratamiento jerarquizado de la misma, en un entorno separado de los sistemas operacionales. (HUMPHRIES et. al, 1999).

Casi todos los conceptos coinciden, por lo que se puede resumir que un DW, es un almacén de datos que es manipulada separadamente de las bases de datos de una organización, la que se obtiene por la integración de información de diversos sistemas de aplicación; u soporta información procesada para proveer una plataforma sólida de datos históricos consolidados para su análisis.

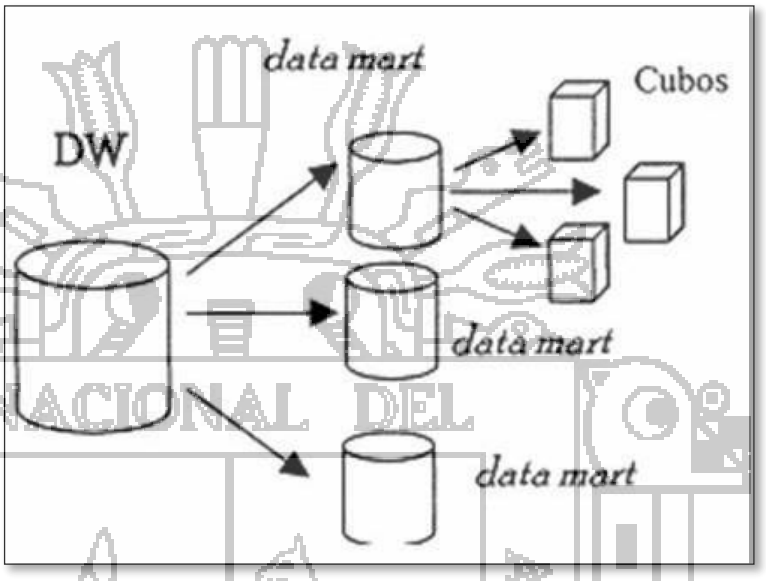
Como se mencionó en el segundo concepto de DW, las características básicas que debe cumplir son: integrado, temático, no volátil e histórico.

#### 2.2.4.1 DataMarts.

Un DW es una agrupación de unidades de información llamados *data marts*. Sin embargo, se considera que un *data mart* es una parte de un DW para un propósito específico (ejemplo, un *data mart* para el departamento de ventas en una empresa).

Un data martes una colección de datos, que es usada para el análisis de consultas dentro de una empresa en uno de sus departamentos o grupo de trabajo (JIAWEI et. Al, 2001).

***Figura N° 4:*** *DW y Data Marts.*




**Fuente:** Presentacion M.Sc. Robert Romero (2011).

### 2.2.3.2 Características de un Data Warehouse.

Un DW tiene varias características que deben considerarse antes de su creación; entre las más importantes se puede mencionar las siguientes: datos integrados, temáticos, históricos y no volátiles. Cada una de estas características se explica en forma detallada a continuación:

### 2.2.4.3 Integrado

Un DW, es construido usualmente por la integración de datos de múltiples fuentes heterogéneas, ya sea desde bases de datos o archivos planos. Las técnicas de limpieza e integración de datos, son aplicadas para garantizar

consistencia en convenciones de nombres uniformes, codificación de estructuras homologadas u atributos de medida, entreoíros (HUMPHRIES et. Al, 1999).

Respecto a las convenciones de nombramiento, uno de los problemas que se presenta más comúnmente en el proceso de integración, es que el mismo elemento es frecuentemente referido por nombres diferentes en las diversas aplicaciones. Los datos almacenados en el DW deben integrarse en una estructura homóloga, por lo que las inconsistencias existentes en los diversos sistemas operacionales, deben ser eliminadas para que al momento de obtener los datos mediante la realización de consultas, los resultados sean confiables.

Otro de los aspectos importantes para lograr la integración de los datos, es traducir las diversas unidades de medida usadas en las diferentes bases de datos, en una medida estándar. Por ejemplo, evitar que en algunas bases de datos, la unidad de medida sean pesos y en otras dólares.

#### 2.2.4.4 Temático

El ambiente operacional se diseña en base a funciones o actividades como: préstamos, facturación, depósitos, etc. Por ejemplo, una aplicación de facturación puede acceder a los datos sobre clientes, productos y precios. La base de datos combina estos elementos en una estructura que acomoda las necesidades de la aplicación.

En el ambiente DW, los datos se organizan en base a sujetos, tales como cliente, vendedor, producto, etc. Por ejemplo, para un fabricante, los sujetos pueden ser clientes, productos, proveedores y vendedores. Para una institución educativa pueden ser estudiantes, clases y profesores. Para un hospital pueden ser pacientes, personal médico, medicamentos, etc.

La diferencia entre la orientación a funciones en los sistemas operacionales y la orientación a sujetos en un DW, radica en el contenido de los datos a nivel de detalle (BERSON et. Al, 1997).

En el DW, se excluye la información que no será usada para la toma de decisiones, mientras que la información utilizada por los sistemas operacionales, contiene datos requeridos para llevar a cabo sus diversas funciones o procesos.

**2.2.4.5 Histórico**

Los datos son almacenados para proveer información desde una perspectiva histórica.

Por lo tanto, cada estructura clave en un DW contiene, implícita o explícitamente, un elemento de tiempo.

En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el presente. Por el contrario, la información de un DW sirve, entre otras cosas, para realizar análisis de tendencias por periodos de tiempo (SALTON, 1993). Por lo tanto, el DW se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.

Como la información en el DW puede ser solicitada en cualquier momento, los datos encontrados en el depósito se llaman de tiempo variante". Los datos históricos son de poco uso en el procesamiento operacional. La información del almacén de datos, por el contrario, debe incluir los datos históricos para usarse en la evaluación de tendencias.

El tiempo variante se muestra de las siguientes formas:

I. En un DW, la información representa los datos sobre un horizonte largo de tiempo (semestres, años). El horizonte de tiempo representado para el ambiente operacional, es mucho más corto (días, semanas).

II. La segunda forma en la que se muestra el tiempo variante en el DW, está en la estructura de sus datos. Cada estructura clave en el DW contiene, implícita o explícitamente, un elemento de tiempo como día, semana, mes, etc. El elemento de tiempo está casi siempre concatenado al dato que lo requiere.

III. La tercera forma en que se detecta el tiempo variante, es cuando la información del DW una vez almacenada correctamente, no puede ser actualizada.

**2.2.3.6 No volátil**

El almacén de información de un DW existe para ser leído, u no modificado. La información es por tanto permanente, por lo que la actualización del DW consiste en la incorporación de los últimos valores que tomaron las distintas variables contenidas, sin realizar ninguna modificación sobre lo que ya existía (HUMPHRIES et. Al, 1999).

La información es útil sólo cuando es estable. Los datos de los Sistemas Operacionales cambian constantemente. Sin embargo, la perspectiva esencial para el análisis en la toma de decisiones, requiere una base de datos estable.

En los sistemas operacionales, la actualización (insertar, borrar u modificar) se hace por registro o por lotes de datos. Pero la manipulación de los datos que ocurre en el DW es mucho más simple, hay dos únicos tipos de operaciones: la carga inicial de datos u el acceso a los mismos. No hay actualización de datos (en el sentido de modificación) en el depósito, como una parte normal de procesamiento, sin embargo se pueden agregar datos nuevos correspondientes al último periodo.

### 2.2.5 OLAP

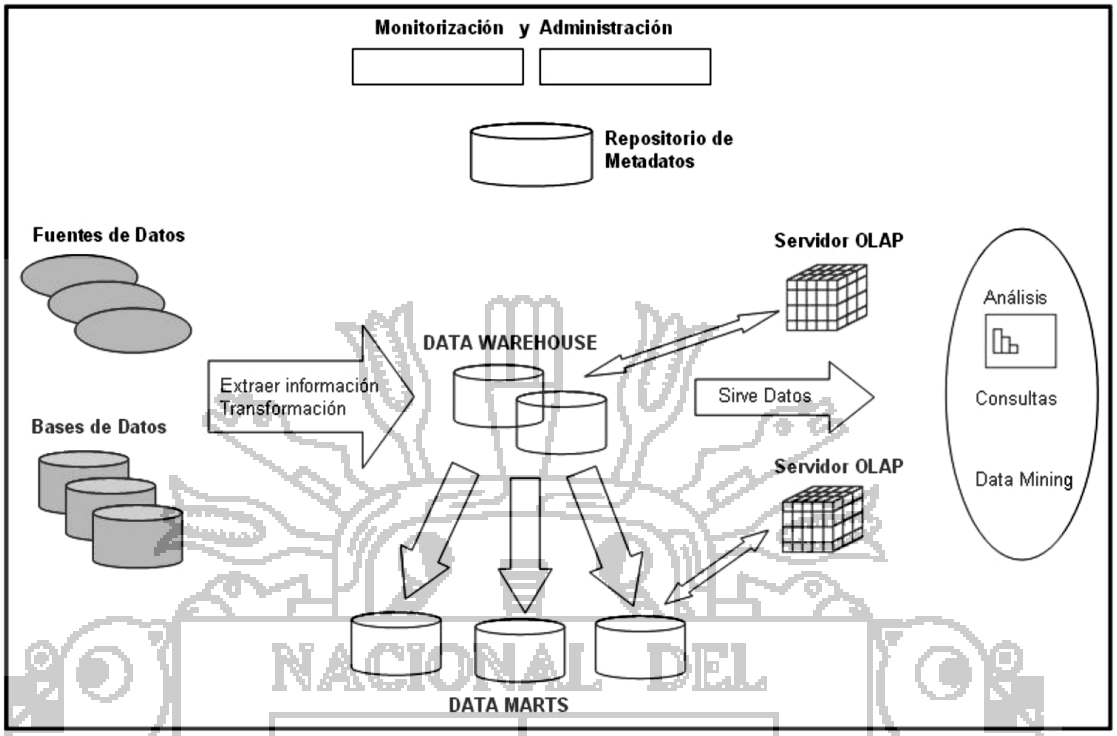
OLAP es una tecnología que ayuda a los trabajadores del conocimiento a hacer con rapidez sus procesos empresariales y la toma de decisiones; permite a analistas y ejecutivos analizar los datos rápidamente, de forma interactiva y teniendo en cuenta varias entidades del negocio (Vassiliadis y Sellis, 1999), (Thomsen, 2002).

Los datos existentes en un sistema DW por lo general se manejan por medio de uno o varios servidores OLAP, como se puede apreciar en la Figura N° 5; esta es una idea que ilustra claramente la fuerte relación entre ambos conceptos.

Estos servidores presentan vistas multidimensionales de los datos a una gran variedad de interfaces: de consulta directa, herramientas para generar informes, herramientas de análisis exploratorio (gráficos, estadística descriptiva, etc.) y herramientas de minería de datos propiamente dichas. Todo esto hace que con los servidores OLAP y sus robustas máquinas de cálculo, los datos históricos almacenados en el DW sean mucho mejor utilizados.

La capacidad que OLAP tiene de poder integrar metadatos con una base de datos relacional, la hace candidata a ser utilizada en sistemas que necesitan extraer los datos de fuentes externas, modelarlos con el uso de metadatos y combinarlos con otros tipos de datos 2.1, para al final obtener un modelo de datos multidimensional. En general, todos los sistemas OLAP cumplen.

***Figura N° 5:*** *Arquitectura del proceso data warehousing.*




Con la capacidad de tratar gran cantidad de datos modelados con un número ilimitado de dimensiones, y sus tiempos de respuesta a consultas muy rápidos los hace una buena opción tecnológica.

Si al proceso de extraer conocimiento de los datos, se le relaciona con técnicas de Minería de Datos y/o Minería de Textos (Delgado et al., 2002), y además se le incluyen procesamientos OLAP, podría lograrse un sistema data warehousing de importantes prestaciones; esta idea forma parte de la motivación de la presente investigación.

### 2.2.6 Base de datos

Es una colección de archivos interrelacionados, son creados con un DBMS, Su contenido engloba la información concerniente de una organización de tal manera que los datos estén disponibles para los usuarios. Su finalidad es eliminar la redundancia o al menos minimizarla.

#### 2.2.5.1 Diseño de la Base de Datos

Santillan Casilla indica que "*En el diseño de la Base de Datos conviene descomponer el proceso del diseño en varias etapas; en cada una se obtiene un resultado intermedio que sirve de punto de partida de la etapa siguiente*". (CASILLAS, Bases de Datos, 2005)

**2.2.5.2 Etapa del Diseño Conceptual:**

La etapa del diseño conceptual nos permite concentrarnos únicamente en la problemática de la estructuración de la información, sin tener que preocuparnos al mismo tiempo de resolver cuestiones tecnológicas.

El resultado de la etapa del diseño conceptual se expresa mediante algún modelo de datos de alto nivel, uno de los más empleados es el modelo entidad – interrelación. (ER).

**2.2.5.3 Etapa del Diseño Lógico:**

Parte de la etapa del diseño conceptual, que se transforma de forma que se adapte a la tecnología que se debe emplear, es preciso que se ajuste al modelo del SGBD con el que se desea implementar la BD. Esta etapa obtendrá un conjunto de relaciones con sus atributos, claves primarias y claves foráneas.

**2.2.5.4 Etapa del Diseño Físico:**

En esta etapa se transforma la estructura obtenida en la etapa de diseño lógico, con el objetivo de conseguir una mayor eficiencia, además, se completa con aspectos de implementación física que dependerán del SGBD. Los aspectos de implementación física que hay que completar consisten normalmente en la elección de las estructuras físicas de implementación de las relaciones.

**2.2.5.5 Normalización de la Base de Datos**

Santilla Casillán menciona que "*La normalización es el proceso mediante el cual se transforman datos complejos a un conjunto de estructuras de datos*

*más pequeñas, que además de ser más simples y más estables, son más fáciles de mantener. También se puede entender la normalización como una serie de reglas que sirven para ayudar a los diseñadores de bases de datos a desarrollar un esquema que minimice los problemas de lógica. Cada regla está basada en la que le antecede. "* (CASILLAS, Bases de Datos, 2005).

### 2.2.5.6 Grados de Normalización

Existen básicamente tres niveles de normalización: Primera Forma Normal (1NF), Segunda Forma Normal (2NF) y Tercera Forma Normal (3NF). Cada una de estas formas tiene sus propias reglas.

Cuando una base de datos se conforma a un nivel, se considera normalizada a esa forma de normalización. No siempre es una buena idea tener una base de datos conformada en el nivel más alto de normalización, puede llevar a un nivel de complejidad que pudiera ser evitado si estuviera en un nivel más bajo de normalización.

#### 2.2.5.6.1 Primera Forma Normal:

La regla de la Primera Forma Normal establece que las columnas repetidas deben eliminarse y colocarse en tablas separadas. Muy a menudo, los diseñadores de bases de datos inexpertos harán algo similar a la tabla no normalizada. Una y otra vez, crearán columnas que representen los mismos datos. Santillán casilla resalta que "*La normalización ayuda a clarificar la base de datos y a organizarla en partes más pequeñas y más fáciles de entender. En lugar de tener que entender una tabla gigantesca y monolítica que tiene muchos diferentes aspectos, sólo tenemos que entender los objetos pequeños y más tangibles, así como las relaciones que guardan con otros objetos también pequeños*" (CASILLAS, Bases de Datos, 2005).

#### 2.2.5.6.2 Segunda Forma Normal:

Santillán Casilla establece que "*La regla de la Segunda Forma Normal establece que todas las dependencias parciales se deben eliminar y separar dentro de sus propias tablas. Una dependencia parcial es un término que describe a aquellos datos que no dependen de la llave primaria de la tabla para identificarlos*." (CASILLAS, Bases de Datos, 2005)

Una vez alcanzado el nivel de la Segunda Forma Normal, se controlan la mayoría de los problemas de lógica. Podemos insertar un registro sin un exceso de datos en la mayoría de las tablas.

#### 2.2.5.6.3 Tercera Forma Normal:

Una tabla está normalizada en esta forma si todas las columnas que no son llave son funcionalmente dependientes por completo de la llave primaria y no hay dependencias transitivas. Santillán Casilla indica que: "*una dependencia transitiva es aquella en la cual existen columnas que no son llave que dependen de otras columnas que tampoco son llave*". (CASILLAS, Bases de Datos, 2005)

En la Tabla N°2 siguiente se describe brevemente en que consiste cada una de las reglas, y posteriormente se explican con más detalle.

***Tabla N° 4:*** *Grados de Normalización*

| Regla | Descripción |
|---|---|
| Primera Forma Normal | Incluye la eliminación de todos los grupos repetidos. |
| Segunda Forma Normal | Asegura que todas las columnas que no son llaves sean completamente dependientes de la llave primaria. |

| | |
|---|---|
| Tercera Forma Normal | Elimina cualquier dependencia transitiva. Una dependencia es aquella en la cual las columnas que no son llave son dependientes de otras columnas que tampoco son llave. |

### 2.2.6 UML (Unified Modeling Language).

Es un lenguaje que permite modelar, construir y documentar los elementos que forman un sistema software orientado a objetos. Se ha convertido en el estándar de facto de la industria, debido a que ha sido concebido por los autores de los tres métodos más usados de orientación a objetos: Grady Booch, Ivar Jacobson y Jim Rumbaugh.

Estos autores fueron contratados por la empresa Rational Software Co. para crear una notación unificada en la que basar la construcción de sus herramientas CASE. En el proceso de creación de UML han participado, no obstante, otras empresas de gran peso en la industria como Microsoft, Hewlett-Packard, Oracle o IBM, así como grupos de analistas y desarrolladores.

Esta notación ha sido ampliamente aceptada debido al prestigio de sus creadores y debido a que incorpora las principales ventajas de cada uno de los métodos particulares en los que se basa: Booch, OMT y OOSE. UML ha puesto fin a las llamadas “guerras de métodos” que se han mantenido a lo largo de los 90, en las que los principales métodos sacaban nuevas versiones que incorporaban las técnicas de los demás. Con UML se fusiona la notación de estas técnicas para formar una herramienta compartida entre todos los ingenieros software que trabajan en el desarrollo orientado a objetos.

#### 2.2.6.1 Diagrama de Casos de Uso.

Un Diagrama de Casos de Uso muestra la relación entre los actores y los casos de uso del sistema. Representa la funcionalidad que ofrece el sistema en lo que se refiere a su interacción externa.

#### 2.2.6.1.1 Elementos.

Los elementos que pueden aparecer en un Diagrama de Casos de Uso son: actores, casos de uso y relaciones entre casos de uso.

#### 2.2.6.1.2 Actores.

Un actor es una entidad externa al sistema que realiza algún tipo de interacción con el mismo. Se representa mediante una figura humana dibujada con palotes. Esta representación sirve tanto para actores que son personas como para otro tipo de actores (otros sistemas, sensores, etc.).

#### 2.2.6.1.3 Casos de Uso.

Un caso de uso es una descripción de la secuencia de interacciones que se producen entre un actor y el sistema, cuando el actor usa el sistema para llevar a cabo una tarea específica. Expresa una unidad coherente de funcionalidad, y se representa en el Diagrama de Casos de Uso mediante una elipse con el nombre del caso de uso en su interior. El nombre del caso de uso debe reflejar la tarea específica que el actor desea llevar a cabo usando el sistema.

#### 2.2.6.1.4 Relaciones entre Casos de Uso.

Entre dos casos de uso puede haber las siguientes relaciones:

a) **Extiende**: Cuando un caso de uso especializa a otro extendiendo su funcionalidad.

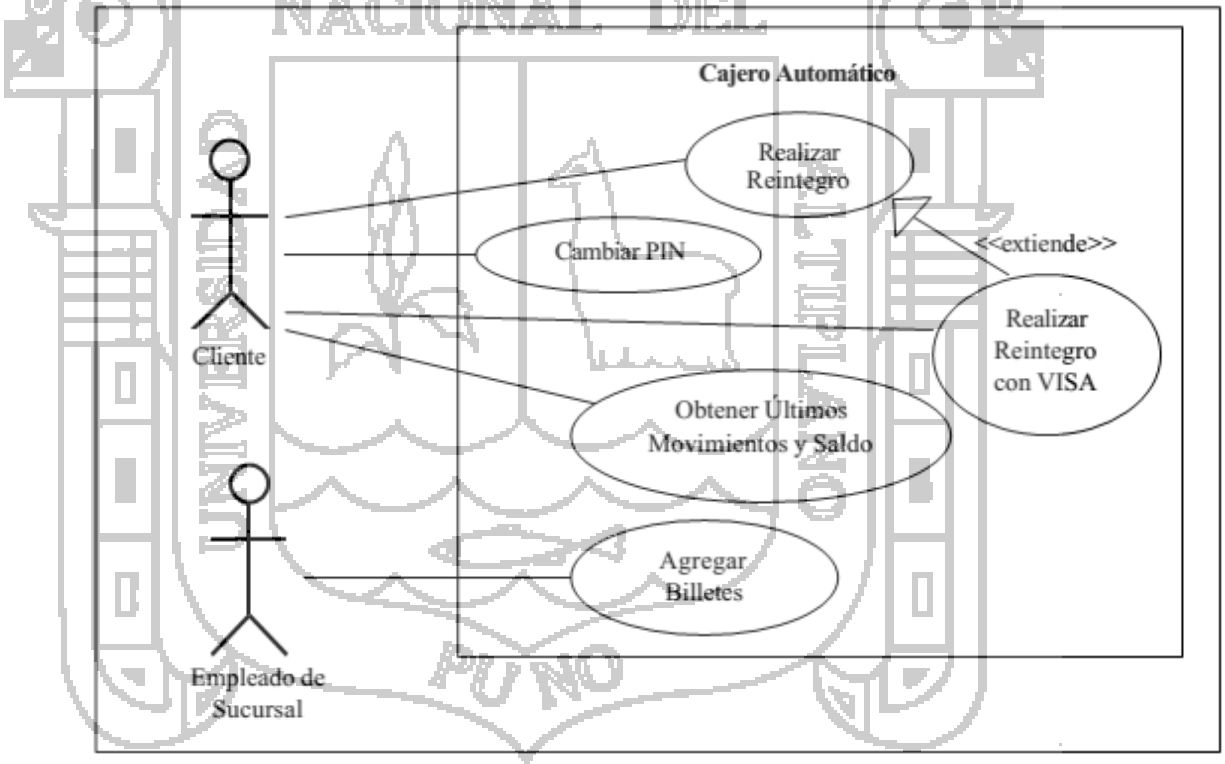
b) **Usa**: Cuando un caso de uso utiliza a otro.

Se representan como una línea que une a los dos casos de uso relacionados, con una flecha en forma de triángulo y con una etiqueta <<extiende>> o <<usa>> según sea el tipo de relación.

En el diagrama de casos de uso se representa también el sistema como una caja rectangular con el nombre en su interior. Los casos de uso están en el interior de la caja del sistema, y los actores fuera, y cada actor está unido a los casos de uso en los que participa mediante una línea.

En la Figura N° 6 se muestra un ejemplo de Diagrama de Casos de Uso para un cajero automático.

***Figura N° 6**: Diagrama de Caso de Uso.*




**Elaboración:** Por los Investigadores.

### 2.2.6.2 Diagramas de Interacción.

En los diagramas de interacción se muestra un patrón de interacción entre objetos. Hay dos tipos de diagrama de interacción, ambos basados en la

misma información, pero cada uno enfatizando un aspecto particular: Diagramas de Secuencia y Diagramas de Colaboración.

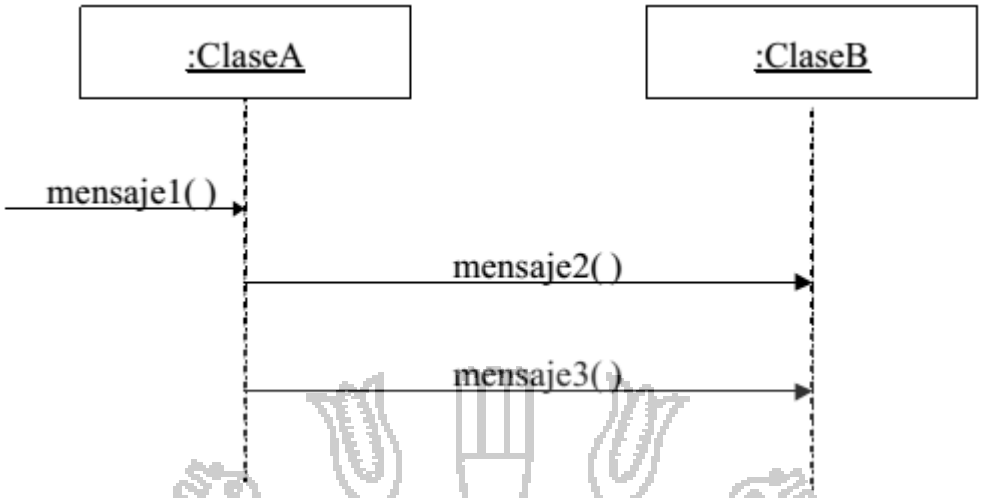
#### 2.2.6.2.1 Diagrama de Secuencia.

Un diagrama de Secuencia muestra una interacción ordenada según la secuencia temporal de eventos. En particular, muestra los objetos participantes en la interacción y los mensajes que intercambian ordenados según su secuencia en el tiempo.

El eje vertical representa el tiempo, y en el eje horizontal se colocan los objetos y actores participantes en la interacción, sin un orden prefijado. Cada objeto o actor tiene una línea vertical, y los mensajes se representan mediante flechas entre los distintos objetos. El tiempo fluye de arriba abajo.

Se pueden colocar etiquetas (como restricciones de tiempo, descripciones de acciones, etc.) bien en el margen izquierdo o bien junto a las transiciones o activaciones a las que se refieren.

***Figura N°* 7:** Diagrama de Secuencia.

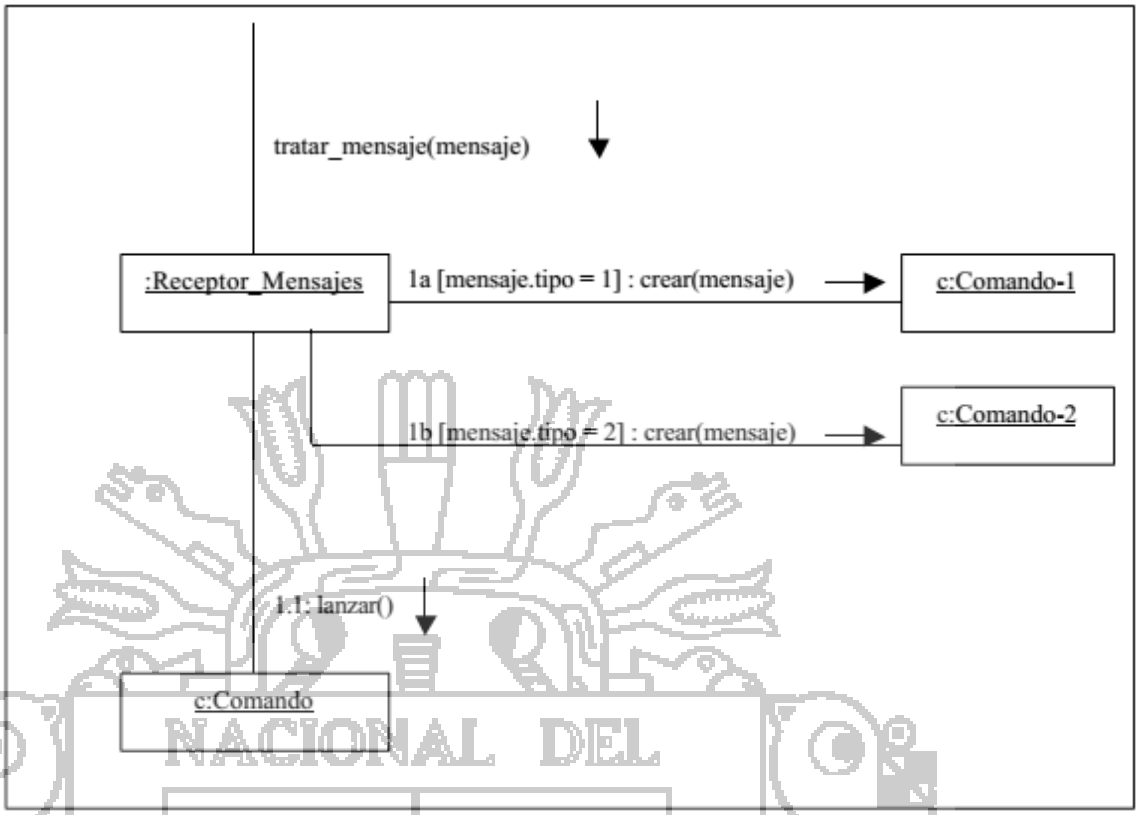



**Elaboración:** Por los investigadores.

#### 2.2.6.2.2 Diagrama de Colaboración.

Un Diagrama de Colaboración muestra una interacción organizada basándose en los objetos que toman parte en la interacción y los enlaces entre los mismos (en cuanto a la interacción se refiere). A diferencia de los Diagramas de Secuencia, los Diagramas de Colaboración muestran las relaciones entre los roles de los objetos. La secuencia de los mensajes y los flujos de ejecución concurrentes deben determinarse explícitamente mediante números de secuencia.

***Figura N° 8:*** *Diagrama de Colaboración.*




**Elaboración:** Por los investigadores.

En cuanto a la representación, un Diagrama de Colaboración muestra a una serie de objetos con los enlaces entre los mismos, y con los mensajes que se intercambian dichos objetos. Los mensajes son flechas que van junto al enlace por el que "circulan", y con el nombre del mensaje y los parámetros (si los tiene) entre paréntesis. Cada mensaje lleva un número de secuencia que denota cuál es el mensaje que le precede, excepto el mensaje que inicia el diagrama, que no lleva número de secuencia. Se pueden indicar alternativas con condiciones entre corchetes (por ejemplo 3 *[condición_de_test] :nombre_de_método()* ), tal y como aparece en el ejemplo de la Figura N° 8. También se puede mostrar el anidamiento de mensajes con números de secuencia como 2.1, que significa que el mensaje con número de secuencia 2no acaba de ejecutarse hasta que no se han ejecutado todos los 2.

**2.2.6.2.3 Diagrama de Estados.**

Un Diagrama de Estados muestra la secuencia de estados por los que pasa un caso de uso o un objeto a lo largo de su vida, indicando qué eventos hacen que se pase de un estado a otro y cuáles son las respuestas y acciones que genera.

En cuanto a la representación, un diagrama de estados es un grafo cuyos nodos son estados y cuyos arcos dirigidos son transiciones etiquetadas con los nombres de los eventos.

Un estado se representa como una caja redondeada con el nombre del estado en su interior. Una transición se representa como una flecha desde el estado origen al estado destino.

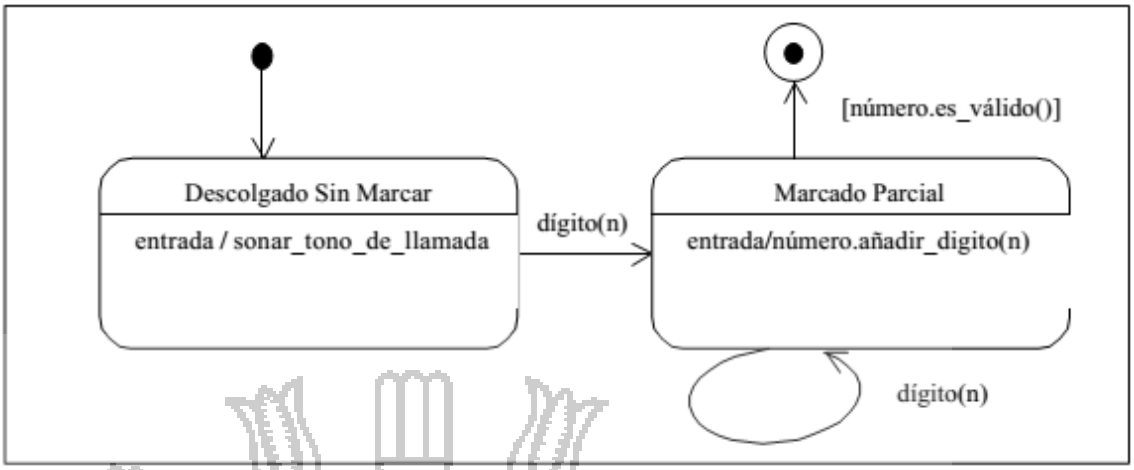
La caja de un estado puede tener 1 o 2 compartimentos. En el primer compartimento aparece el nombre del estado. El segundo compartimento es opcional, y en él pueden aparecer acciones de entrada, de salida y acciones internas.

Una acción de entrada aparece en la forma *entrada/acción_asociada* donde *acción_asociada* es el nombre de la acción que se realiza al entrar en ese estado. Cada vez que se entra al estado por medio de una transición la acción de entrada se ejecuta.

Una acción de salida aparece en la forma *salida/acción_asociada*. Cada vez que se sale del estado por una transición de salida la acción de salida se ejecuta.

Una acción interna es una acción que se ejecuta cuando se recibe un determinado evento en ese estado, pero que no causa una transición a otro estado. Se indica en la forma *nombre_de_evento/acción_asociada*.

***Figura N° 9:*** *Diagrama de Estados.*



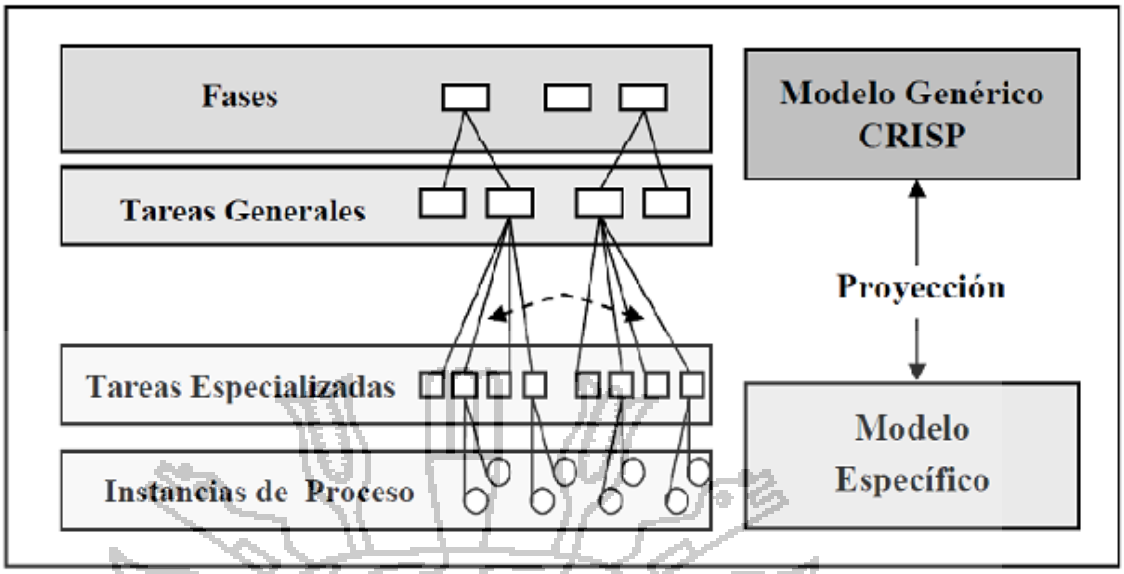

**Elaboración:** Por los investigadores.

Un diagrama de estados puede representar ciclos continuos o bien una vida finita, en la que hay un estado inicial de creación y un estado final de destrucción (del caso de uso o del objeto). El estado inicial se muestra como un círculo sólido y el estado final como un círculo sólido rodeado de otro círculo. En realidad, los estados inicial y final son pseudoestados, pues un objeto no puede "estar" en esos estados, pero nos sirven para saber cuáles son la transición inicial y final(es).

### 2.2.7 Metodología CRISP-DM.

Cross Industry Standard Process for Data Mining, es la guía de referencia más ampliamente utilizada en el desarrollo de proyectos de Data Mining. Está dividida en 4 niveles de abstracción organizados de forma jerárquica (Figura N° 10) en tareas que van desde el nivel más general, hasta los casos más específicos y organiza el desarrollo de un proyecto de Data Mining, en una serie de seis fases (Figura N° 10).

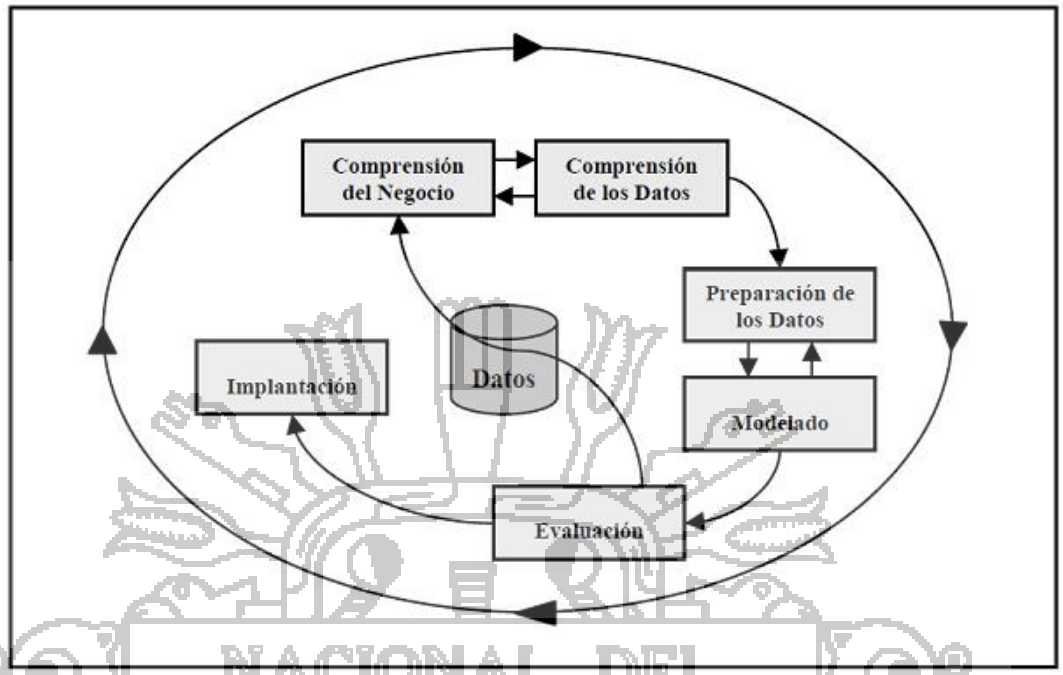
***Figura N° 10**: Esquema de 4 niveles de CRISP-DM*




**Fuente:** Metodología CRISP-DM.

La sucesión de fases no es necesariamente rígida. Cada fase es estructurada en varias tareas generales de segundo nivel. Las tareas generales se proyectan a tareas específicas, donde finalmente se describen las acciones que deben ser desarrolladas para situaciones específicas, pero en ningún momento se propone como realizarlas.

***Figura N° 11**: Modelo de Proceso CRISP-DM.*




**Fuente:** Metodología CRISP-DM.

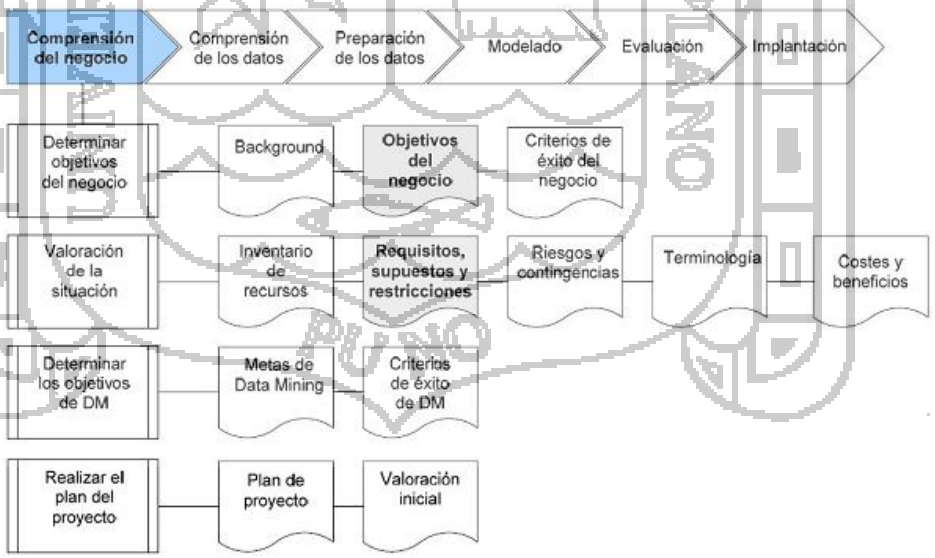
### 2.2.10.1 Fase de comprensión del negocio o problema

La primera fase de la guía de referencia CRISP-DM, denominada fase de comprensión del negocio o problema (Figura N° 8), es probablemente la más importante y aglutina las tareas de comprensión de los objetivos y requisitos del proyecto desde una perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto. Sin lograr comprender dichos objetivos, ningún algoritmo por muy sofisticado que sea, permitirá obtener resultados fiables. Para obtener el mejor provecho de Data Mining, es necesario entender de la manera más completa el problema que se desea resolver, esto permitirá recolectar los datos correctos e interpretar correctamente los resultados. En esta fase, es muy importante la capacidad de poder convertir el conocimiento adquirido del negocio, en un problema de Data Mining y en un plan preliminar cuya

meta sea el alcanzar los objetivos del negocio. Una descripción de cada una de las principales tareas que componen esta fase es la siguiente:

*Determinar los objetivos del negocio*. Esta es la primera tarea a desarrollar y tiene como metas, determinar cuál es el problema que se desea resolver, por qué la necesidad de utilizar Data Mining y definir los criterios de éxito. Los problemas pueden ser diversos como por ejemplo, detectar fraude en el uso de tarjetas de crédito, detección de intentos de ingreso indebido a un sistema, asegurar el éxito de una determinada campaña publicitaria, etc. En cuanto a los criterios de éxito, estos pueden ser de tipo cualitativo, en cuyo caso un experto en el área de dominio, califica el resultado del proceso de DM, o de tipo cuantitativo, por ejemplo, el número de detecciones de fraude o la respuesta de clientes ante una campaña publicitaria.

***Figura N° 12**: Fase de comprensión del negocio.*




**Fuente:** Metodología CRISP-DM.

*Evaluación de la situación*. En esta tarea se debe calificar el estado de la situación antes de iniciar el proceso de DM, considerando

aspectos tales como: ¿cuál es el conocimiento previo disponible acerca del problema?, ¿se cuenta con la cantidad de datos requerida para resolver el problema?, ¿cuál es la relación coste beneficio de la aplicación de DM?, etc. En esta fase se definen los requisitos del problema, tanto en términos de negocio como en términos de Data Mining.

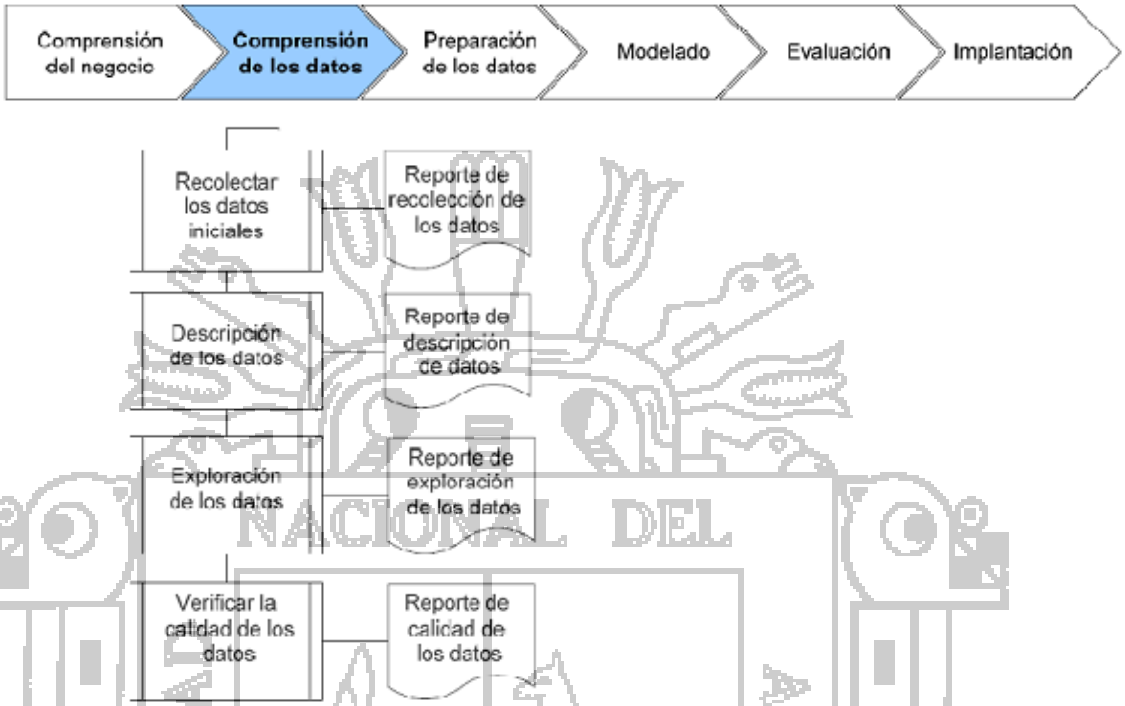
*Determinación de los objetivos de DM*. Esta tarea tiene como objetivo representar los objetivos del negocio en términos de las metas del proyecto de DM, como por ejemplo, si el objetivo del negocio es el desarrollo de una campaña publicitaria para incrementar la asignación de créditos hipotecarios, la meta de DM será por ejemplo, determinar el perfil de los clientes respecto de su capacidad de endeudamiento. Producción de un plan del proyecto. Finalmente esta última tarea de la primera fase de CRISP-DM, tiene como meta desarrollar un plan para el proyecto, que describa los pasos a seguir y las técnicas a emplear en cada paso.

**2.2.10.2 Fase de la compresión de los datos**

La segunda fase (Figura N° 13), fase de comprensión de los datos, comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las próximas dos fases, son las que demandan el mayor esfuerzo y tiempo en un proyecto de DM. Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos ad-hoc al proyecto de DM, pues durante el desarrollo del proyecto, es posible que se generen frecuentes y

abundantes accesos a la base de datos a objeto de realizar consultas y probablemente modificaciones, lo cual podría generar muchos problemas.

***Figura N° 13**: Fase de compresión de los datos.*




**Fuente:** Metodología CRISP-DM.

Las principales tareas a desarrollar en esta fase del proceso son:

*Recolección de datos iniciales*. La primera tarea en esta segunda fase del proceso de CRISP-DM, es la recolección de los datos iniciales y su adecuación para el futuro procesamiento. Esta tarea tiene como objetivo, elaborar informes con una lista de los datos adquiridos, su localización, las técnicas utilizadas en su recolección y los problemas y soluciones inherentes a este proceso.

*Descripción de los datos*. Después de adquiridos los datos iniciales, estos deben ser descritos.

Este proceso involucra establecer volúmenes de datos (número de registros y campos por registro), su identificación, el significado de cada campo y la descripción del formato inicial.

*Exploración de datos*. A continuación, se procede a su exploración, cuyo fin es encontrar una estructura general para los datos. Esto involucra la aplicación de pruebas estadísticas básicas, que revelen propiedades en los datos recién adquiridos, se crean tablas de frecuencia y se construyen gráficos de distribución. La salida de esta tarea es un informe de exploración de los datos.

*Verificación de la calidad de los datos*. En esta tarea, se efectúan verificaciones sobre los datos, para determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los valores nulos, y para encontrar valores fuera de rango, los cuales pueden constituirse en ruido para el proceso. La idea en este punto, es asegurar la completitud y corrección de los datos.

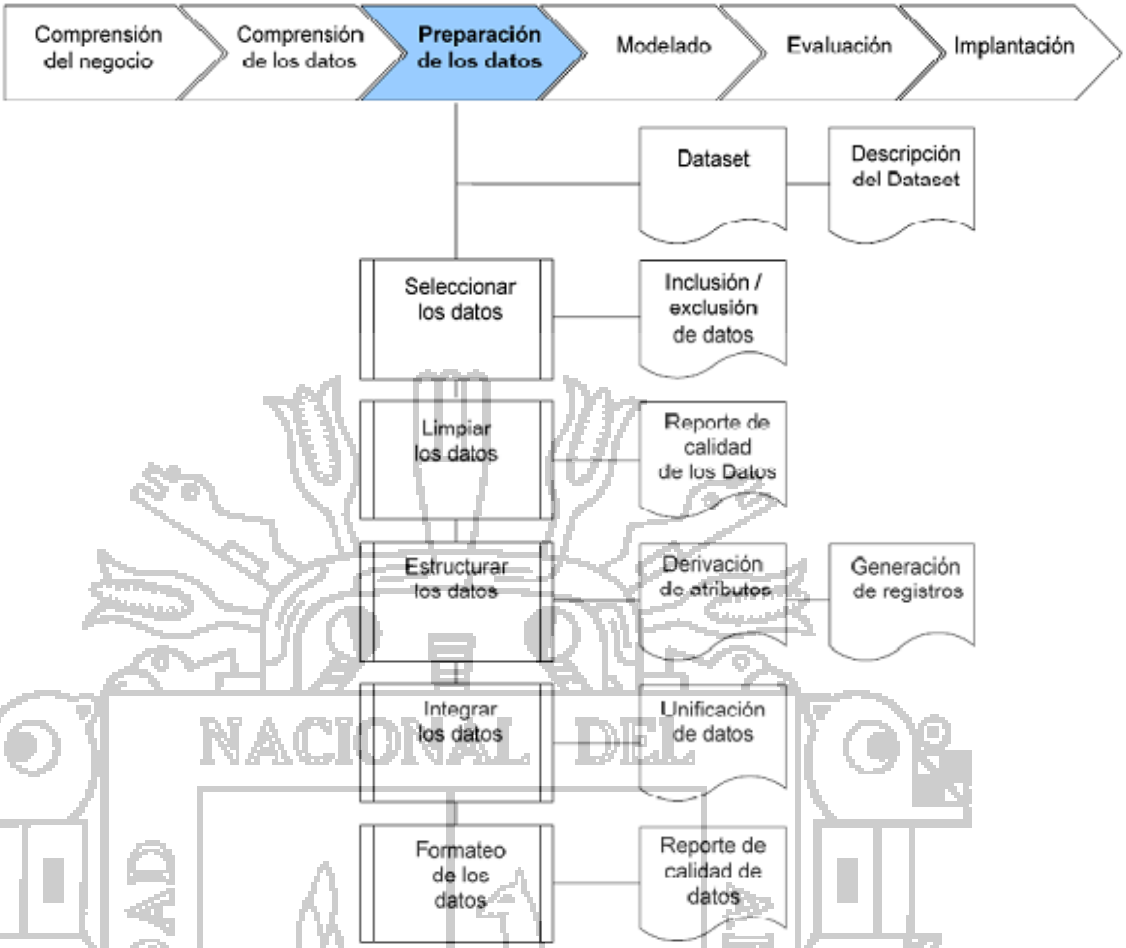
**2.2.10.3 Fase de preparación de los datos**

En esta fase y una vez efectuada la recolección inicial de datos, se procede a su preparación para adaptarlos a las técnicas de Data Mining que se utilicen posteriormente, tales como técnicas de visualización de datos, de búsqueda de relaciones entre variables u otras medidas para exploración de los datos. La preparación de datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

Esta fase se encuentra relacionada con la fase de modelado, puesto que en función de la técnica de modelado elegida, los datos requieren ser procesados de diferentes formas. Es así que las fases de preparación y modelado interactúan de forma permanente. La Figura N° 14, ilustra las áreas de que se compone ésta, e identifica sus salidas. Una descripción de las tareas involucradas en esta fase es la siguiente: *Selección de datos*. En esta etapa, se selecciona un subconjunto de los datos adquiridos en la fase anterior, apoyándose en criterios previamente establecidos en las fases anteriores: calidad de los datos en cuanto a completitud y corrección de los datos y limitaciones en el volumen o en los tipos de datos que están relacionadas con las técnicas de DM seleccionadas. *Limpieza de los datos*. Esta tarea complementa a la anterior, y es una de las que más tiempo y esfuerzo consume, debido a la diversidad de técnicas que pueden aplicarse para optimizar la calidad de los datos a objeto de prepararlos para la fase de modelación. Algunas de las técnicas a utilizar para este propósito son: normalización de los datos, discretización de campos numéricos, tratamiento de valores ausentes, reducción del volumen de datos, etc.

*Estructuración de los datos*. Esta tarea incluye las operaciones de preparación de los datos tales como la generación de nuevos atributos a partir de atributos ya existentes, integración de nuevos registros o transformación de valores para atributos existentes.

***Figura N° 14**: Fase de preparación de los datos*




**Fuente:** Metodología CRISP-DM.

*Integración de los datos*. La integración de los datos, involucra la creación de nuevas estructuras, a partir de los datos seleccionados, por ejemplo, generación de nuevos campos a partir de otros existentes, creación de nuevos registros, fusión de tablas campos o nuevas tablas donde se resumen características de múltiples registros o de otros campos en nuevas tablas de resumen.

*Formateo de los datos*. Esta tarea consiste principalmente, en la realización de transformaciones sintácticas de los datos sin modificar su significado, esto, con la idea de permitir o facilitar el empleo de alguna técnica de DM en particular, como por ejemplo la reordenación de los campos y/o registros de la tabla o el ajuste de los valores de los campos

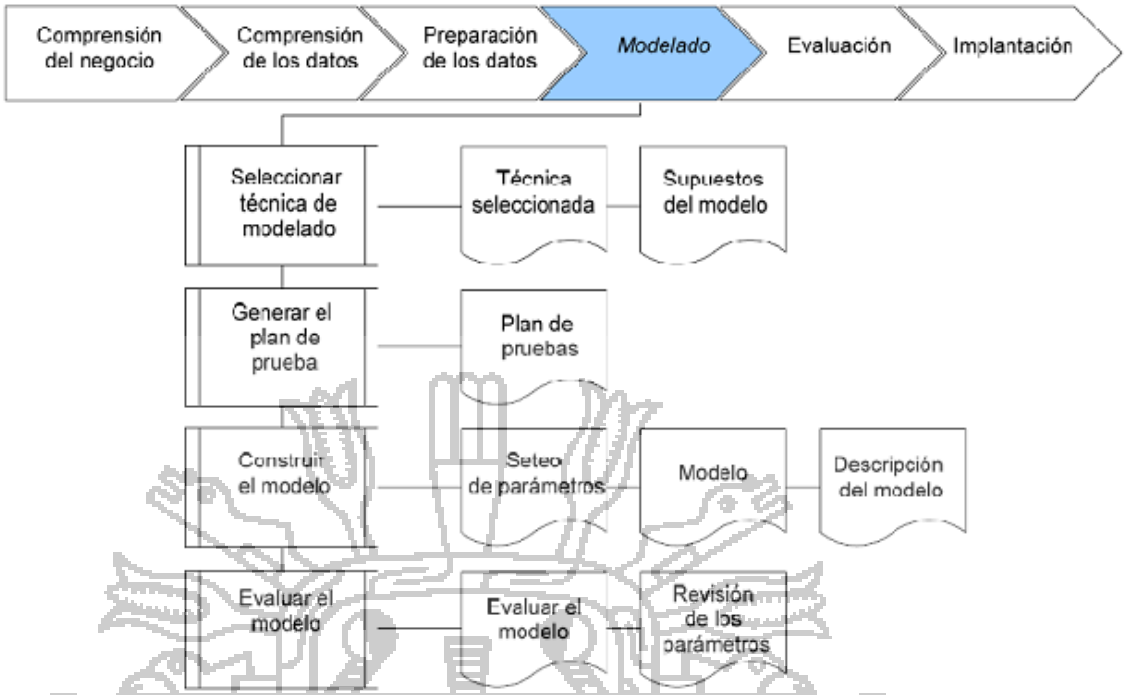
a las limitaciones de las herramientas de modelación (eliminar comas, tabuladores, caracteres especiales, máximos y mínimos para las cadenas de caracteres, etc.).

#### 2.2.10.4 Fase de Modelado

En esta fase de CRISP-DM, se seleccionan las técnicas de modelado más apropiadas para el proyecto de Data Mining específico. Las técnicas a utilizar en esta fase se eligen en función de los siguientes criterios:

- Ser apropiada al problema.
- Disponer de datos adecuados.
- Cumplir los requisitos del problema.
- Tiempo adecuado para obtener un modelo.
- Conocimiento de la técnica.

***Figura N° 15**: Fase de modelado*




**Fuente:** Metodología CRISP-DM.

Previamente al modelado de los datos, se debe determinar un método de evaluación de los modelos que permita establecer el grado de bondad de ellos. Después de concluir estas tareas genéricas, se procede a la generación y evaluación del modelo. Los parámetros utilizados en la generación del modelo, dependen de las características de los datos y de las características de precisión que se quieran lograr con el modelo. La Figura N° 15 ilustra las tareas y resultados que se obtienen en esta fase. Una descripción de las principales tareas de esta fase es la siguiente:

*Selección de la técnica de modelado*. Esta tarea consiste en la selección de la técnica de DM más apropiada al tipo de problema a resolver. Para esta selección, se debe considerar el objetivo principal del proyecto y la relación con las herramientas de DM existentes. Por ejemplo, si el problema es de clasificación, se podrá elegir de entre árboles de decisión, k-nearest neighbour o razonamiento basado en casos (CBR); si el problema

es de predicción, análisis de regresión, redes neuronales; o si el problema es de segmentación, redes neuronales, técnicas de visualización, etc.

*Generación del plan de prueba*. Una vez construido un modelo, se debe generar un procedimiento destinado a probar la calidad y validez del mismo. Por ejemplo, en una tarea supervisada de DM como la clasificación, es común usar la razón de error como medida de la calidad. Entonces, típicamente se separan los datos en dos conjuntos, uno de entrenamiento y otro de prueba, para luego construir el modelo basado en el conjunto de entrenamiento y medir la calidad del modelo generado con el conjunto de prueba.

*Construcción del Modelo*. Después de seleccionada la técnica, se ejecuta sobre los datos previamente preparados para generar uno o más modelos. Todas las técnicas de modelado tienen un conjunto de parámetros que determinan las características del modelo a generar. La selección de los mejores parámetros es un proceso iterativo y se basa exclusivamente en los resultados generados. Estos deben ser interpretados y su rendimiento justificado.

*Evaluación del modelo*. En esta tarea, los ingenieros de DM interpretan los modelos de acuerdo al conocimiento preexistente del dominio y los criterios de éxito preestablecidos. Expertos en el dominio del problema juzgan los modelos dentro del contexto del dominio y expertos en Data Mining aplican sus propios criterios (seguridad del conjunto de prueba, perdida o ganancia de tablas, etc...).

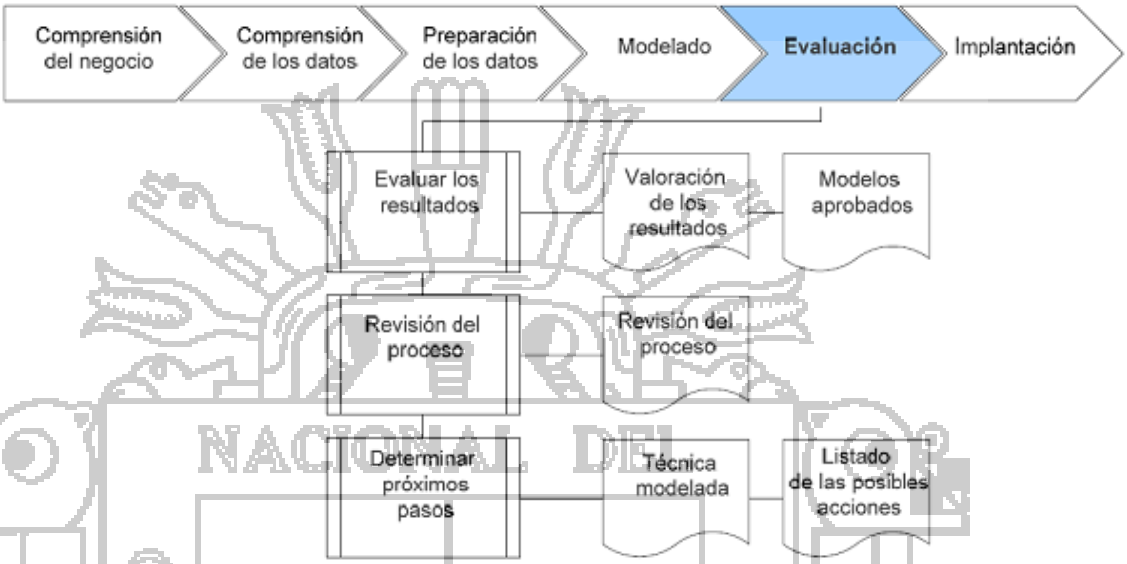
#### 2.2.10.5 Fase de evaluación

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse además, que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se haya posiblemente cometido algún error. Considerar que se pueden emplear múltiples herramientas para la interpretación de los resultados. *Las matrices de confusión* (Edelstein, 1999) son muy empleadas en problemas de clasificación y consisten en una tabla que indica cuantas clasificaciones se han hecho para cada tipo, la diagonal de la tabla representa las clasificaciones correctas. Si el modelo generado es válido en función de los criterios de éxito establecidos en la fase anterior, se procede a la explotación del modelo. La Figura N° 16 detalla las tareas que componen esta fase y los resultados que se deben obtener. Las tareas involucradas en esta fase del proceso son las siguientes:

*Evaluación de los resultados*. En los pasos de evaluación anteriores, se trataron factores tales como la exactitud y generalidad del modelo generado. Esta tarea involucra la evaluación del modelo en relación a los objetivos del negocio y busca determinar si hay alguna razón de negocio para la cual, el modelo sea deficiente, o si es aconsejable probar el modelo, en un problema real si el tiempo y restricciones lo permiten. Además de los resultados directamente relacionados con el objetivo del proyecto, ¿es aconsejable evaluar el modelo en relación a otros objetivos distintos a los originales?, esto podría revelar información adicional.

*Proceso de revisión*. El proceso de revisión, se refiere a calificar al proceso entero de DM, a objeto de identificar elementos que pudieran ser mejorados.

***Figura N° 16**: Fase de Evaluación.*




**Fuente:** Metodología CRISP-DM.

*Determinación de futuras fases*. Si se ha determinado que las fases hasta este momento han generado resultados satisfactorios, podría pasarse a la fase siguiente, en caso contrario podría decidirse por otra iteración desde la fase de preparación de datos o de modelación con otros parámetros. Podría ser incluso que en esta fase se decida partir desde cero con un nuevo proyecto de DM.

### 2.2.10.6 Fase de Implementación

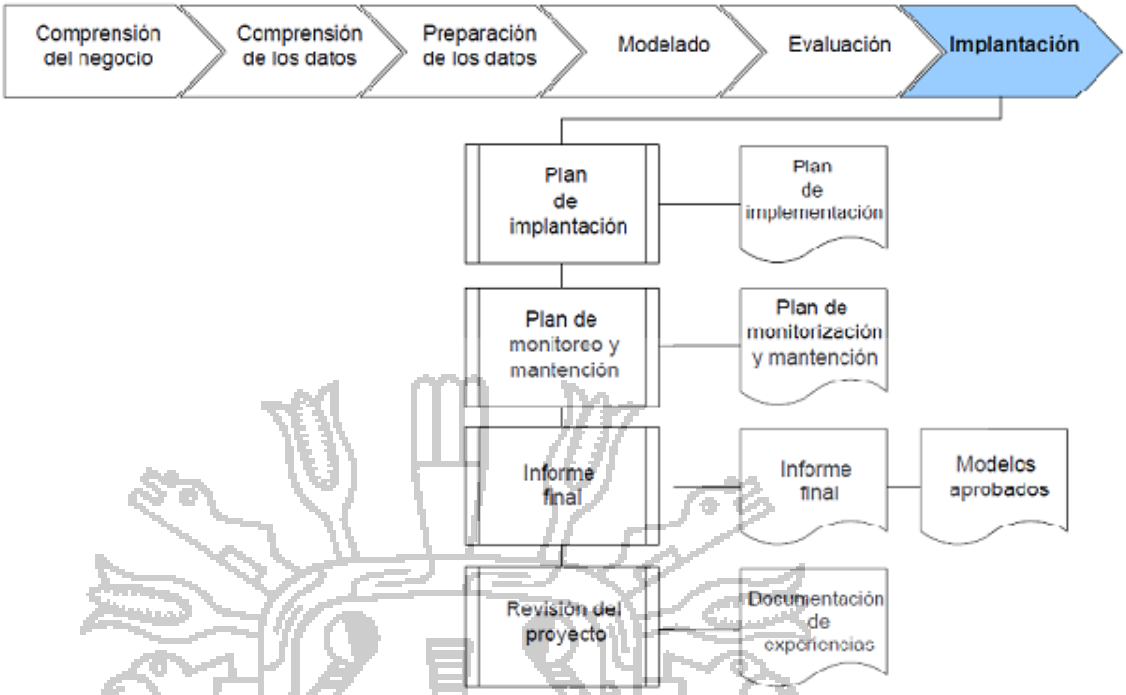
En esta fase (Figura N° 17), y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones

basadas en la observación del modelo y sus resultados, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso, como por ejemplo, en análisis de riesgo crediticio, detección de fraudes, etc. Generalmente un proyecto de Data Mining no concluye en la implantación del modelo, pues se deben documentar y presentar los resultados de manera comprensible para el usuario, con el objetivo de lograr un incremento del conocimiento.

Por otra parte, en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados. Las tareas que se ejecutan en esta fase son las siguientes:

*Plan de implementación*. Para implementar el resultado de DM en la organización, esta tarea toma los resultados de la evaluación y concluye una estrategia para su implementación. Si un procedimiento general se ha identificado para crear el modelo, este procedimiento debe ser documentado para su posterior implementación. Monitorización y Mantenimiento. Si los modelos resultantes del proceso de Data Mining son implementados en el dominio del problema como parte de la rutina diaria, es aconsejable preparar estrategias de monitorización y mantenimiento para ser aplicadas sobre los modelos. La retroalimentación generada por la monitorización y mantenimiento pueden indicar si el modelo está siendo utilizado apropiadamente.

***Figura N° 17**: Fase de Implementación.*




**Fuente:** Metodología CRISP-DM.

*Informe Final*. Es la conclusión del proyecto de DM realizado. Dependiendo del plan de implementación, este informe puede ser sólo un resumen de los puntos importantes del proyecto y la experiencia lograda o puede ser una presentación final que incluya y explique los resultados logrados con el proyecto. Revisión del proyecto: En este punto se evalúa qué fue lo correcto y qué lo incorrecto, qué es lo que se hizo bien y qué es lo que se requiere mejorar.

### 2.2.8 Gestión de Portafolio Crediticio – Cosechas Crediticias

Es la manipulación y/o transformación de datos del portafolio de créditos en la rama de Cosechas Crediticias, de manera que genere conocimiento relevante para la toma de decisiones, el proceso de gestión sigue fases de descubrimiento de conocimiento.

Las cosechas crediticias es un componente de la Gestión del Portafolio Crediticio, que representa la generación de conocimiento por medio de reportes

que son analizados por un analista de riesgos u otro ejecutivo que interpreta el conocimiento. Las cosechas crediticias miden la calidad de colocación y el periodo de maduración de un crédito.

#### 2.2.7.1 Pre Procesado de Datos.

Dada la base de datos de créditos con los datos ya procesados, tal como se muestra en la tabla:

***Tabla N° 5***: *Campos de la Base de Datos, Información Procesada.*

| CABECERA | DESCRIPCIÓN |
|---|---|
| Fecha Desembolso | Es la fecha en el que el cliente desembolso del crédito otorgado. |
| Fecha de Cierre | Es la fecha transcurrida mayor a la fecha desembolso, generalmente en intervalos de 30 días, esta fecha puede ser del último periodo cerrado o también denominada fecha actual. |
| Soc | Id que identifica al cliente. |
| Monto Desembolsado | Monto pactado entre el cliente y la empresa financiera para beneficio del primero. Este valor no variara. |
| Saldo Capital | Monto restante por pagar en los plazos de pago pactados. Esto plazos determinan la cuota de pago. |
| Saldo Vencido | Monto restante por pagar debido al atraso de pago de la última o últimas cuota(s) no pagada(s) en el plazo pactado. Si no existe ninguna cuota atrasada a la Fecha de Cierre este valor es 0 (cero). |
| Producto | Denominación del producto crediticio que se le otorgo al cliente. |
| Oficina | Denominación de la Agencia en donde se otorgó el crédito al cliente. Y donde se realizó la operación crediticia. |

| | |
|---|---|
| Asesor | Oficial de crédito encargado del asesoramiento de crédito del cliente. Es quien capto al cliente o también puede ser heredado. |
| Modalidad de Crédito | Es el número de días como intervalos de plazo, puede ser de 30, 60, 90, 120, y 180 días. |
| Crédito Castigado | Crédito irrecuperable el cual ha sido dado en castigo, que ya no se cobra al cliente, donde el Saldo Capital y el Saldo Vencido son el mismo, y se mantendrán en las futuras fechas de cierre. Este valor es binominal SI y NO. |
| Crédito Campaña | Crédito desembolsado bajo un beneficio para el cliente. Este valor es binominal SI y NO. |

**Fuente:** Caja Los Andes.

**Elaboración:** Por los Investigadores.

**a) Consideraciones Generales de los Datos:**

**i. Dinámica de Fecha de Cierre.**

Considerando que cada 30 días o fines de mes, se consolida una base de datos de los clientes, de los cuales se duplicaran algunos campos, sin embargo el campo Fecha de Cierre tendrá un valor diferente, esto hace que algunos campos se repitan en la base de datos.

Si un crédito es nuevo, el campo Fecha de Desembolso y el campo Fecha de Cierre, tendrán los mismos valores.

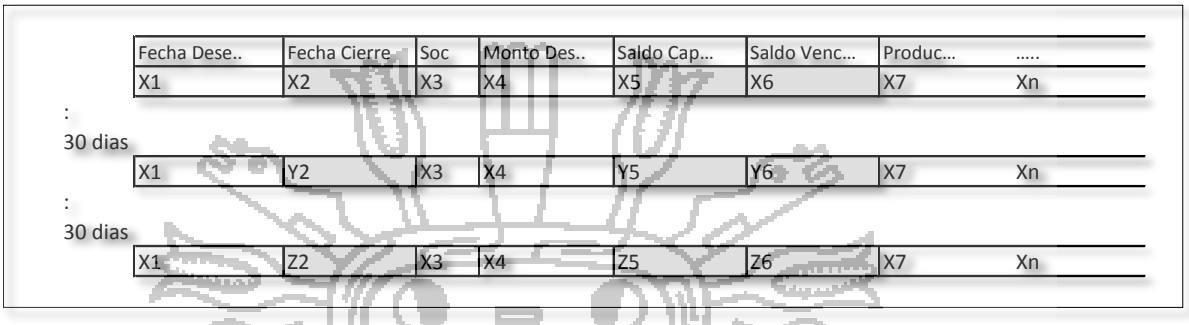
**ii. Dinámica de Saldo Capital y Saldo Vencido.**

Pasado el número de días dados en el plazo, este siempre estará contemplado en una fecha, que será la Fecha de Cierre, debido a que los plazos están dados por los múltiplos de 30, bajo un supuesto canónico cumplido el plazo de la cuota el cliente efectúa el pago, por tanto esto disminuirá el Saldo Capital

restante por pagar, y este será un valor diferente al que ya se tenía en la base de datos. Si no se efectúa el pago de cuota pactada entonces el Saldo Vencido tendrá el valor de Saldo Capital.

***Figura N° 18:*** *Dinámica de Datos por el transcurso de tiempo.*

**b) Agrupación de Datos.**

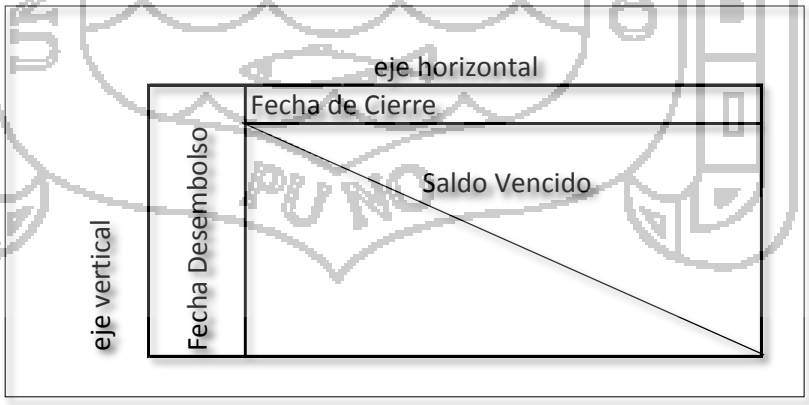



**Elaboración:** Por los investigadores.

La agrupación está dado por dos ejes, en el eje vertical se tiene el campo Fecha de Desembolso, y en el eje horizontal el campo Fecha de Cierre. Los datos agrupados esta dados por los comprendidos en el campo Saldo Capital, Saldo Vencido y Monto Desembolsado. Filtrados por los restantes datos.

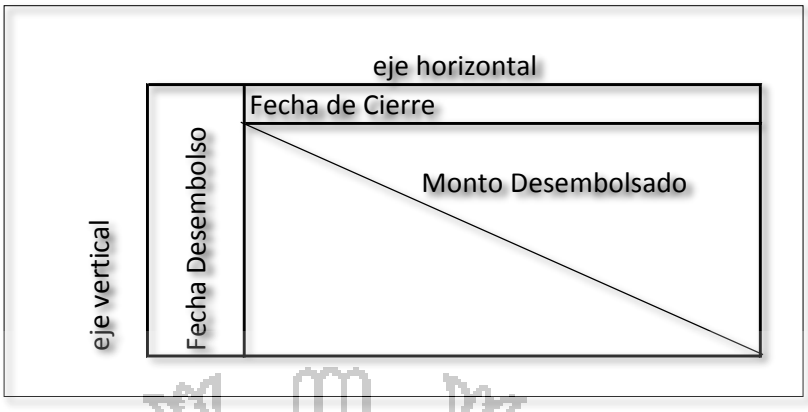
**i. Agrupación Saldo Vencido.**

***Figura N° 19:*** *Agrupación a Saldo Vencido.*




**ii. Agrupación a Monto Desembolsado.**

***Figura N° 20:*** *Agrupación a Monto Desembolsado*




**Elaboración:** Por los investigadores.

### 2.2.7.2 Presentación.

Para la presentación de la información primero se tendrá que calcular el Índice de Fallidos.

**a) Fallido**.

Es el valor porcentual que resulta de la relación del valor de Saldo Vencido y Monto Desembolsado, de una misma fecha de desembolso y fecha de cierre.

$$Fallido_{ik} = \frac{Saldo\ Vencido_{\{(Fecha\ de\ Desembolso_i)|(Fecha\ de\ Cierre_k)\}}}{Monto\ Desembolsado_{(Fecha\ de\ Desembolso_i)}}$$

Es índice da a conocer que porcentaje del Monto Total desembolsado en la fecha "i" de desembolso, que entro en vencimiento; debemos considerar que Saldo Vencido refiere a los pagos incumplidos de los montos colocados en la fecha "i".

La presentación de los fallidos como está dada bajo la misma agrupación anterior.

***Tabla N° 6:*** *Presentación de Fallidos.*

Eje Horizontal, Fecha de Cierre

| Eje Vertical, Fecha de Desembol | X1 | X2 | X3 | X4 | X5 | ... | ... | Xn |
|---|---|---|---|---|---|---|---|---|
| X1 | SV1,1 / MD1 | SV1,2 / MD1 | SV1,3 / MD1 | SV1,4 / MD1 | SV1,5 / MD1 | | | SV1,n / MD1 |
| X2 | | SV2,2 / MD2 | SV2,3 / MD2 | SV2,4 / MD2 | SV2,5 / MD2 | | | SV2,n / MD2 |
| X3 | | | SV3,3 / MD3 | SV3,4 / MD3 | SV3,5 / MD3 | | | SV3,n / MD3 |
| X4 | | | | SV4,4 / MD4 | SV4,5 / MD4 | | | SV4,n / MD4 |
| X5 | | | | | SV5,5 / MD5 | | | SV5,n / MD5 |
| : | | | | | | ... | ... | : |
| : | | | | | | | ... | : |
| Xn | | | | | | | | SVn,n / MDn |

**Elaboración:** Por los investigadores.

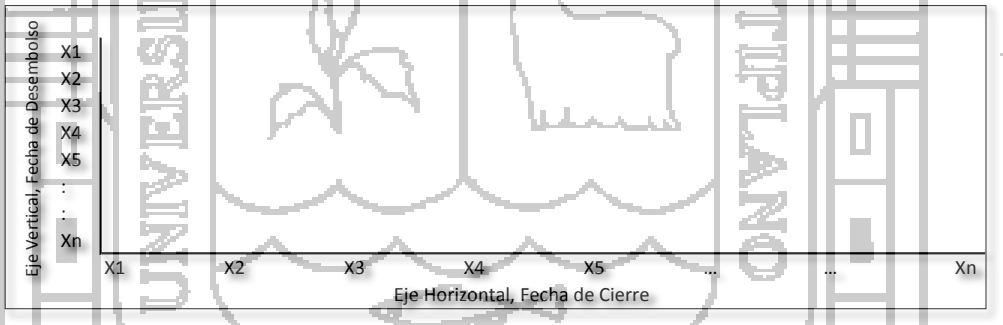
### 2.2.7.3 Interpretación del Conocimiento.

Dada la presentación, como instancia final de la gestión de portafolio es la interpretación del conocimiento de la cual deriva la toma de decisiones.

La interpretación del conocimiento se por la observación del comportamiento del índice fallido y se determina de tres maneras.

**a) Comportamiento Canónico.**

***Grafico N° 1**: Comportamiento Canónico de Fallidos*

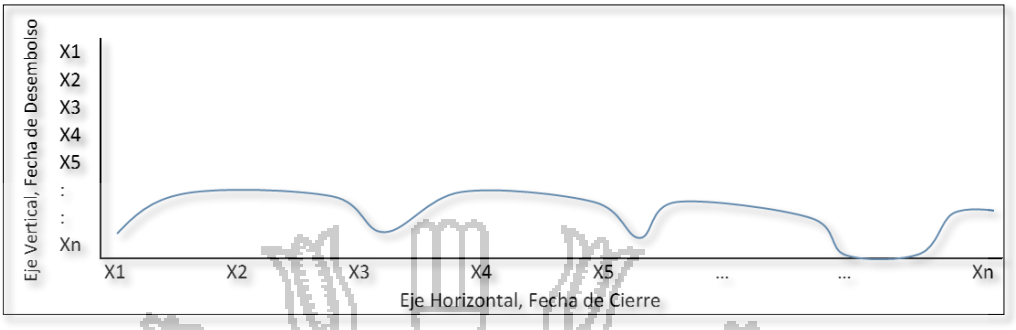



**Elaboración:** Por los investigadores.

Sucede cuando los valores de Saldo Vencido son cero (0), lo que significa que el crédito está siguiendo el normal proceso de pago por parte del cliente.

**b) Comportamiento Estable o de Estabilidad.**

***Grafico N° 2:*** *Comportamiento Estable de Fallidos.*

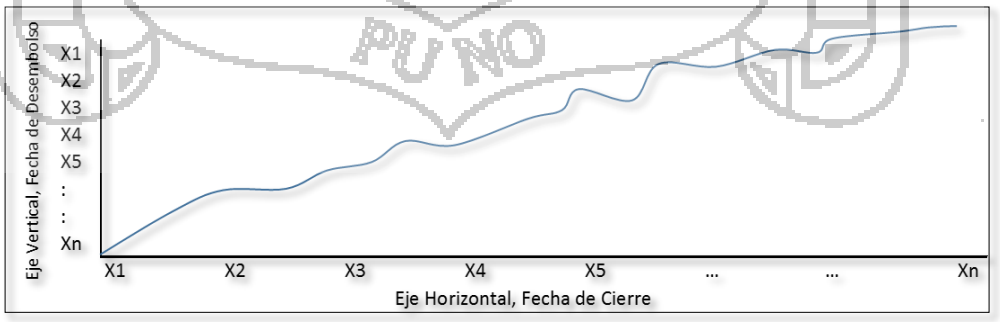



**Elaboración:** Por los investigadores.

Dado el incumplimiento del deudor en alguna cuota pactada hasta la fecha de cierre, el valor de Saldo Capital se adopta a Saldo Vencido, por lo que dada la relación a Monto Desembolsado, tendrá como resultado que el índice de fallido posea valores. Sin embargo estos valores pueden permanecer hasta que el cliente cumpla con el pago respectivo, y automáticamente el Saldo Vencido se convierte en Cero, y dicho pago disminuirá el Saldo Capital, este comportamiento de vaivén del pago hace que el comportamiento del índice fallido se estabilice o se mantenga.

**c) Comportamiento No Estable.**

***Grafico N° 3:*** *Comportamiento No Estable de fallidos.*




**Elaboración:** Por los investigadores.

Este comportamiento se define como el no deseable, pues representa un incremento de Saldo Vencido, en forma incontrolada; y refleja incumplimientos de pagos en muchos clientes; podría significar una pérdida.

### 2.2.7 Métricas de Desarrollo de Software.

El autor Norman define las métricas de software como: "la aplicación continua de mediciones basadas en técnicas para el proceso de desarrollo del software y sus productos para suministrar información relevante a tiempo, así el administrador junto con el empleo de estas técnicas mejorará el proceso y sus productos." (NORMAN & LAWRENCE, 1997).

Las métricas del software responden a dos objetivos que David Card menciona: "valorar y estimar las magnitudes objeto de valoración son tres: la calidad, fiabilidad productividad. La estimación parte de mediciones históricas para prever el esfuerzo y el tiempo que debe invertirse en un proyecto dado, y las características del resultado final." (CARD & GLASS, 1990).

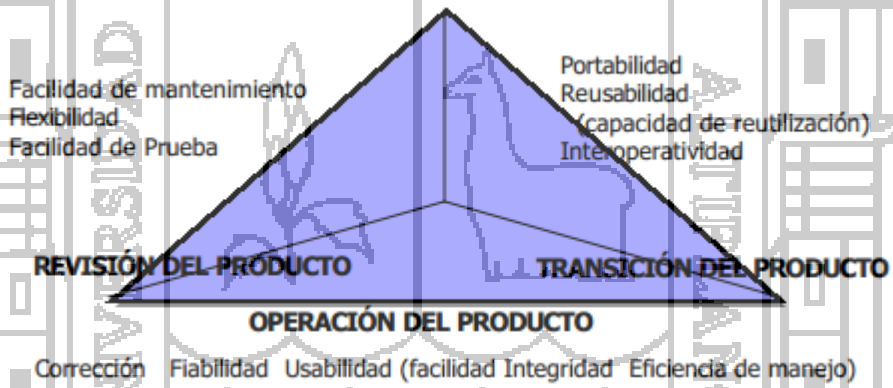
El autor Vegan Lebrún menciona que: "hay muchas magnitudes que pueden ser medidas en el software: el tamaño en líneas de código, el coste monetario del desarrollo, el tiempo de desarrollo en días de trabajo, el tamaño de la memoria precisada en bytes, e incluso el número de quejas del usuario antes de entregar el producto. Diferentes observadores del mismo producto, pueden obtener distintas medidas, incluso en una misma magnitud." (LEBRÚN & SANTILLÁN, 2008).

En las secciones siguientes detallaremos algunas métricas aplicables al proceso de desarrollo del software, es decir a los requerimientos, análisis, diseño, implementación y pruebas.

### 2.2.8.1 Grado de satisfacción de requerimientos

La medición de la satisfacción de los requerimientos de usuario, se mide con las métricas de calidad del producto, para lo cual McCall y sus colegas plantearon una categorización de factores que afectan a la calidad de software, en donde se centralizan con tres aspectos importantes de un producto software que lo referencia Presman: "características operativas, capacidad de cambio y adaptabilidad a nuevos entornos, refiriéndose a los factores McCall proporciona las siguientes descripciones:" (PRESMAN, 2002).

***Figura N° 21:*** *Factores que afectan la calidad de Software. (Fuente: Pressman, 2002).*




El cálculo de cada factor de calidad se realizó con la fórmula de McCall:

$$F_q = (C_1 x m_1) + (C_2 x m_2) + .. + (C_n x m_n)$$

Donde:
$F_q$: es un factor de calidad,
$C_n$: Coeficiente de regresión
$m_n$: Métricas que afectan al factor de calidad.

La relación entre los factores de calidad de software y las métricas de la Tabla 4, se tiene que ponderar el peso para cada métrica dependiendo de los productos y negocios locales.

***Tabla N° 7:*** *Factores de calidad y Métricas de calidad de Software.*

| Métrica de la calidad de software / Factor de calidad | Corrección | Fiabilidad | Eficiencia | Integridad | Mantenimiento | Flexibilidad | Capacidad de | Portabilidad | Reusabilidad | Interoperabilidad | Usabilidad |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Facilidad de auditoria | | | x | | | | x | | | | |
| Exactitud | | x | | | | | | | | | |
| Estandarización de Comunicaciones | | | | | | | | | | x | |
| Compleción | | x | | | x | x | | | | | |
| Complejidad | x | | | | | | | | | | |
| Concisión | x | x | | | | x | x | | | | |
| Consistencia | | | | x | x | x | | | | | |
| Estandarización de datos | x | x | | | | x | x | | | x | |
| Tolerancia a errores | | x | | | | | | | | | |
| Eficiencia de ejecución | | | x | | | | | | | | |
| Capacidad de expansión | | | | | | x | | | | | |
| Generalidad | | | | | | x | | | x | x | x |
| Independencia del hardware | | | | | | | | x | x | | |
| Instrumentación | | | | | x | x | x | | | | |
| Modularidad | x | | | | | x | x | x | x | x | |
| Operatividad | | | x | | | | | | | | x |
| Seguridad | | | x | | | | | | | | |
| Auto documentación | | | | | x | | | | | | |
| Simplicidad | | | | | x | x | x | | x | x | |
| Independencia del sistema | | x | | | | | x | x | x | | |
| Trazabilidad | | | | | | | | x | x | | |
| Facilidad de formación | | | | | | | | | | | x |

### 2.2.9 Métricas de puntos de función

La métrica de puntos de función es una métrica que se puede aplicar en las primeras fases de desarrollo. El autor Lawrence se refiere a los puntos de función como: “una métrica para establecer el tamaño y complejidad software en base a la cantidad de funcionalidad requerida y entregada a los usuarios, o una función

que mide el tamaño lógico o funcional de los proyectos." (NORMAN & LAWRENCE, 1997).

Por su parte Juan Busquelle menciona que: "el análisis de los Puntos de Función es la medida del tamaño de las funciones de usuario, de la aplicación o de parte de ella. Las funciones de usuario son los componentes solicitados y reconocidos por el usuario, que se toman de las especificaciones que describen lo que el software debe hacer para satisfacer las necesidades del mismo." (BUSQUELLE, 2010).

Según Lebrúm se determinan cinco características de dominios de información las cuáles son: el número de entradas, salidas y peticiones del usuario; el número de archivos e interfaces externas; además menciona que: "A los datos de la tabla se les asocia un valor de complejidad y obtener una cuenta total, que vienen a ser el Total de puntos de función no ajustados." (LEBRÚN & SANTILLÁN, 2008).

Santillán describe el procedimiento para calcular el factor de ajuste para el cálculo de los puntos de función y recomienda que: "Los 14 factores que muestra la Tabla 5, son características generales y se deben analizar, evitando adivinar características, para ello siempre que sea necesario, hay que conversar con el usuario principal del proyecto o del área en la cual la duda se relacione, éstos asignan a cada uno de los 14 parámetros un valor de 0 a 5 según la influencia del mismo en el proyecto." (LEBRÚN & SANTILLÁN, 2008).

***Tabla N° 8:*** *Resumen de las características generales del sistema (Fuente: Lebrun y Santillan, 2008)*

| **Características** | **Influencia** |
|---|---|
| Comunicación de datos | [0-5] |
| Procesamiento distribuido | [0-5] |
| Desempeño | [0-5] |

| | |
|---|---|
| ConFigura N°ción del equipamiento | [0-5] |
| Volumen de transacciones | [0-5] |
| Entrada de datos online | [0-5] |
| Procesamiento complejo | [0-5] |
| Reusabilidad | [0-5] |
| Facilidad de implementación | [0-5] |
| Facilidad de operación | [0-5] |
| Múltiples locales | [0-5] |
| Facilidad de cambios | [0-5] |
| **Nivel de influencia** | NI |

**Elaboración:** Por los investigadores.

El valor de ajuste se obtiene a partir del nivel de influencia, y con esto el punto de función con ajuste.

### 2.2.10 Métrica de complejidad del sistema

La métrica de complejidad del sistema; complejidad estructural y complejidad de datos; es una métrica para medir el diseño arquitectónico.

Roger Presman menciona que: "las métricas de diseño de alto nivel se concentran en las características de la arquitectura del programa con especial énfasis en la estructura arquitectónica y en la eficiencia de los módulos." (PRESMAN, 2002).

Los autores Card y Glass definen tres medidas de la complejidad del diseño del software: la complejidad estructural, la complejidad de datos y complejidad del sistema.

La complejidad estructural, $S(i)$, de un módulo $i$ se define de la siguiente manera: $S(i) = f^2_{out}(i)$, donde $f^2_{out}(i)$, es la expansión del módulo $i$. La complejidad de datos, $D(i)$, proporciona una indicación de la complejidad en la interfaz interna de un módulo i y s e define como:

$$D_{(i)} = \frac{V(i)}{f_{out}(i) + 1}$$

Dónde *v(i)*, es el número de variables de entrada y salida que entran y salen del módulo i. Finalmente la complejidad del sistema, *C(i)*, se define como la suma de las complejidades estructural y de datos.

Card y Glass mencionan que: “A medida que crecen los valores de complejidad, la complejidad arquitectónica o global del sistema también aumenta llevando a una mayor probabilidad de que aumente el esfuerzo necesario para la integración y las pruebas.” (CARD & GLASS, 1990).

### 2.2.11 Métricas de código fuente

Presman referencia a Halstead y menciona que: “la teoría de la ciencia del software propuesta por Halstead es probablemente la medida de complejidad mejor conocida y minuciosamente estudiada. La ciencia del software propuso la primera ley analítica y cuantitativa para el software de computadora.” (PRESMAN, 2002).

Roger Presman menciona también que: “las métricas de Halstead utiliza un conjunto de medidas primitivas que puede obtenerse una vez que se han generado o estimado el código después de completar el diseño. Halstead utiliza medidas primitivas para desarrollar expresiones para la longitud global del programa, volumen mínimo potencial para un algoritmo; el volumen real, número de bits requeridos para especificar un programa; el nivel del programa, una medida de complejidad del software, nivel del lenguaje, una constante para un lenguaje dado.” (PRESMAN, 2002).

### 2.2.12 Métricas de pruebas

El autor Roger Presman menciona que: “la mayoría de las métricas para pruebas se concentran en el proceso de prueba, no en las características técnicas de las pruebas mismas. En general, los responsables de las pruebas

deben fiarse en las métricas de análisis, diseño y código para que sirvan de guía en el diseño y ejecución de los casos de prueba.” (PRESMAN, 2002).

Para medir el esfuerzo de las pruebas Roger sugiere que: “también se puede estimar utilizando métricas obtenidas de las medidas de Halstead. Usando la definición del volumen de un programa, V, y nivel de programa, NP.” (PRESMAN, 2002).

### 2.2.9 Business Process Management (BPM).

Es un conjunto de métodos, herramientas y tecnologías utilizados para diseñar, representar, analizar y controlar procesos de negocio operacionales. BPM es un enfoque centrado en los procesos para mejorar el rendimiento que combina las tecnologías de la información con metodologías de proceso y gobierno. BPM es una colaboración entre personas de negocio y tecnólogos para fomentar procesos de negocio efectivos, ágiles y transparentes.

BPM abarca personas, sistemas, funciones, negocios, clientes, proveedores y socios. Como mucha gente, puede que encuentre este concepto algo confuso. ¿Qué son “procesos de negocio operacionales”? O ¿qué es un enfoque “centrado en los procesos”? ¿Y desde cuándo “colaboran” las personas de negocio con las de tecnología? No se preocupe, vamos a explicarlo todo.

BPM combina métodos ya probados y establecidos de gestión de procesos con una nueva clase de herramientas de software empresarial. Ha posibilitado adelantos muy importantes en cuanta la velocidad y agilidad con que las organizaciones mejoran el rendimiento de negocio. Con BPM:

a) Los directores de negocio pueden, de forma más directa, medir, controlar y responder a todos los aspectos y elementos de sus procesos operacionales.

b) Los directores de tecnologías de la información pueden aplicar sus habilidades y recursos de forma más directa en las operaciones de negocio.

c) La dirección y los empleados de la organización pueden alinear mejor sus esfuerzos y mejorar la productividad y el rendimiento personal.

d) La empresa, como un todo, puede responder de forma más rápida a cambios y desafíos a la hora de cumplir sus fines y objetivos.

e) ¿Demasiado bueno para ser verdad? Pues esta vez lo es. BPM está cambiando rápidamente el panorama de los negocios a escala mundial.

#### 2.2.9.1 BPMN (BUSINESS PROCESS MANAGEMENT NOTATION)

Acrónimo de Business Process Modeling Notation (notación de creación de modelos de procesos de negocio), se trata de una notación gráfica estandarizada para representar los procesos de negocio en un flujo de trabajo, que facilita la mejora de la comunicación y la portabilidad de los modelos de proceso.

## 2.3 DEFINICIÓN DE TÉRMINOS BÁSICOS

### 2.3.1 Sistema de Gestión:

Es una herramienta informática que determinados usuarios especializados pueden utilizar accediendo a un software que automatiza con precisión; procesos de gestión.

### 2.3.2 OLAP:

OLAP es una tecnología que ayuda a los trabajadores del conocimiento a hacer con rapidez sus procesos empresariales y la toma de decisiones; permite a analistas y ejecutivos analizar los datos rápidamente, de forma interactiva y

teniendo en cuenta varias entidades del negocio (Vassiliadis y Sellis, 1999), (Thomsen, 2002).

### 2.3.3 RUP (Proceso Unificado de Rational):

Es una metodología de desarrollo de software que intenta integrar todos los aspectos a tener en cuenta durante todo el ciclo de vida del software, con el objetivo de hacer abarcables tanto pequeños como grandes proyectos software.

### 2.3.4 Data Mining:

Es el proceso analítico, por medio del cual se extrae información oculta de grandes cantidades de datos siendo muy útil para predecir futuros comportamientos y tendencias.

### 2.3.5 Gestión de Portafolio:

Conjunto de procesos para agrupar, presentar e interpretar información creada a partir del portafolio de créditos. Con la intención de hallar índices que muestran tendencias positivas o negativas, que servirán para la toma de decisiones.

## 2.4 OPERACIONALIZACIÓN DE VARIABLES

### 2.4.1 Variable Independiente

Sistema de Gestión de Portafolio Crediticio.

### 2.4.2 Variable Dependiente

Gestión de Portafolio de Créditos.

***Tabla N° 9**: Cuadro de Operacionalización de Variable*

| VARIABLE | DIMENSIONES | INDICADORES | INSTRUMENTO | ESCALA |
|---|---|---|---|---|
| **Variable Independiente**<br><br>Sistema de Gestión | ➢ Determinación de los objetivos. | - Tiempo de Procesamiento<br>- Numero de procesos optimizados. | | |

| | | | | |
|---|---|---|---|---|
| | ➢ Pre Procesamiento de Datos<br>➢ Determinación del Modelo.<br>➢ Análisis de los Resultados | - Modelamiento del negocio.<br>- Generación de Conocimiento.<br>- Funcionalidad | Cuestionario y Contraste Pre y Post | Muy bueno<br>Bueno<br>Regular<br>Malo<br>Muy malo |
| **Variable Dependiente**<br><br>Gestión de Portafolio Crediticio. | ➢ Preparación de Datos.<br>➢ Presentación de los datos procesados<br>➢ Interpretación del Conocimiento. | - Agrupación de datos por ejes.<br>- Resultados de Consultas de Agrupación.<br>- Muestreo de Agrupación.<br>- Resultados optimizados de Cálculo por agrupación.<br>- Resultados optimizados de Programación de cálculo.<br>- Resultado de querys a cálculos.<br>- Muestreo por Interfaces.<br>- Existencia de resultado de cálculos a visualizar.<br>- Existencia de interfaces interpretativos.<br>- Resultados de consultas por interface de usuario. | | Porcentaje de Efectividad |

**Elaboración:** Por los investigadores.

# CAPITULO III

# DISEÑO METODOLÓGICO DE LA INVESTIGACIÓN

## 3.1 DISEÑO DE LA INVESTIGACIÓN

De acuerdo a Hernández "*los diseños cuasi experimentales manipulan deliberadamente al menos una variable independiente para observar su efecto y relación con una o más variables dependientes, sólo que difieren de los experimentos en el grado de confiabilidad que se pueda tener sobre la equivalencia inicial en los grupos, puesto que son grupos intactos*." Hernández (2003).

El diseño de investigación corresponde al tipo **cuasi experimental,** ya que se trabaja con grupos establecidos, el esquema es el siguiente:

***Tabla N° 10:*** *Diseño De Investigación Cuasiexperimental*

| | Pre-Prueba | Tratamiento | Post - Prueba |
|---|---|---|---|
| Grupo Experimental (GE) | $O_1$ | X | $O_2$ |

**Fuente:** Hernández (2003)
**Elaboración:** Los Ejecutores.

**Donde:**

$O_1$ : Diagnostico mediante pruebas sobre el proceso actual de gestión de portafolio crediticio.

X: Sistema de Gestión utilizando tecnología Data Mining.

$O_2$: Pruebas sobre funcionalidad optimización del proceso de gestión de portafolio crediticio utilizando el sistema de gestión utilizando tecnología Data Mining.

## 3.2 POBLACIÓN Y MUESTRA

### 3.2.1 Población de la Investigación

La población está constituido por todos los registros almacenados en los periodos de 2009 hasta Junio del 2014, para el carácter de prueba el volumen

de muestra estará conformado por los registros ingresados desde 1 de Abril del 2012, hasta 31 de Junio del 2014.

***Tabla N° 11**: Periodos de Población.*

| Empresa | Población | | | Total |
|---|---|---|---|---|
| | Fecha Inicio | Fecha Fin | Número de Registros (promedio) | |
| Caja Los Andes | 01/01/2008 | 30/06/2014 | 600,000. | 5.5 años |

**Fuente**: Caja Los Andes, Cartera de Créditos.
**Elaboración**: Los ejecutores.

### 3.2.2 Muestra de la Investigación

La muestra está constituido por los periodos de 2012, 2013 y hasta la fecha de Junio del 2014.

***Tabla N° 12**: Periodos de Muestra.*

| Oficinas Especiales | Muestra | | | Total |
|---|---|---|---|---|
| | Fecha Inicio | Fecha Fin | Número de Registros (promedio) | |
| Ilave | 01/04/2012 | 30/06/2014 | 90 mil. | 2 años |
| Puno | 01/04/2012 | 30/06/2014 | 90 mil | 2 años |
| Juliaca | 01/04/2012 | 30/06/2014 | 90 mil | 2 años |

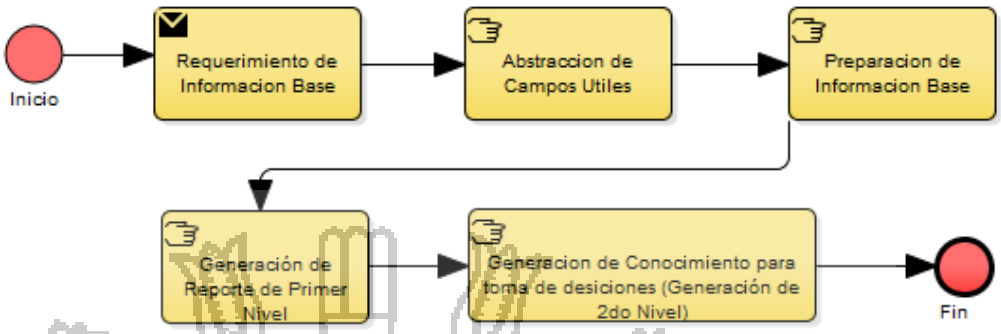
**Elaboración**: Por los Investigadores.

## 3.3 TÉCNICAS E INSTRUMENTOS PARA RECOLECTAR LOS DATOS

### 3.4.8 Análisis de Contenido.

Durante el funcionamiento del modelo de negocio se vino almacenado un buró de información sobre las colocaciones de créditos, donde se registra la información de crédito y del cliente. Para la elaboración de la Reporte de las Cosechas Crediticias como parte de la Gestión de Portafolio Crediticio, se realiza el siguiente flujo de procesos:

***Figura N° 22**: Flujo del proceso pre SGP de Gestión de Portafolio Crediticio – Generación de Cosechas Crediticias*




**Elaboración:** Por los investigadores.

Entonces se identifican que los procesos a mejorar son los que se muestran en la Figura N° 13, y se detalla su estado de elaboración y tiempo.

***Tabla N° 13**: Procesos para la elaboración de Reportes de Cosechas, como parte a optimizar de la Gestión de Portafolio Crediticio.*

| Proceso | Manera | Tiempo |
|---|---|---|
| Requerimiento de Información Base | Correo, Manual | 1H a 3H |
| Abstracción de Campos Útiles de la Información Base | Manual | 2H |
| Preparación de Información base | Manual | 5H |
| Modelado de la Información. | Manual | 2H |
| Generación de Reporte (Conocimiento) | Manual | 2H |
| **TOTAL PROCESOS** | | **TOTAL TIEMPO** |
| **5** | | **14H** |

**Elaboración:** Por los investigadores.

La diagramación de los procesos con la notación BPM (Business Process Management) nos permite identificar como resultado los procesos que se realizan Pre SGP. Entonces para la investigación se constituye como metodología para tratamiento de los datos, que dado el caso son el número de procesos que se deberán optimizar.

### 3.4.9 Observación.

Nótese que en la Figura N° 19, los procesos son manuales, recalcando una vez más el planteamiento del problema. Se aplica mediante la observación de la identificación de datos clave que muestren las mejoras graduales sobre el proceso, para al final demostrarlos en la discusión de resultados.

## 3.4 MÉTODOS DE TRATAMIENTO DE DATOS

La Caja Rural de Ahorro y Crédito Los Andes cuenta con un procedimiento

El tratamiento de datos se realizará tomando en cuenta el siguiente procedimiento:

- Recolección y Evaluación mediante Pruebas y observación.
- Contraste en resultados de las pruebas.
- Análisis de resultados.

### 3.4.1 Material experimental

Los materiales y herramientas que se utilizarán son las siguientes:

- Pentaho.
- Workbench.
- Servidor Apache Tomcat.
- Enterprise Architect.

- Sintaxis MDX
- Syco Analitics.
- Pentaho server BI.

# CAPITULO IV

# ANÁLISIS E INTERPRETACIÓN DE RESULTADOS DE LA INVESTIGACIÓN

## 4.1 Caso de Estudio

La cede corporativa de la Caja Rural de Ahorro y Crédito los Andes, en donde se realizó el caso de estudio se encuentra ubicada en la ciudad de Puno, Jr. Junín Nº 129 viene colocando créditos de manera que estos sirvan de apoyo para el desarrollo del emprendedor rural, ofreciendo sus diferentes productos evaluados acorde a la región. Enfrentando riesgos tales como Crediticio, Liquidez y Mercado, Operacional; como los más comunes.

La Oficina de Riesgos tiene como misión mitigar estos riesgos que a posterior cada riesgo deriva a generar pérdidas financieras, denotando que el riesgo de más perjuicio es el Crediticio que infiere directamente en las perdidas al significar el incumplimiento de pago por parte del deudor.

En los últimos años se ha observado que el 23% de perdidas debido al incumplimiento de las responsabilidades del deudor, significándose esto un riesgo latente el cual debe ser mitigado.

La SBS en la Resolución 3780-2008 plantea que cada Institución Financiera implementara un Sistema de “Cosechas”, que es parte del Sistema de Portafolio crediticio, y también se constituye como parte del proceso de Gestión de Portafolio.

Para poder realizar el análisis es necesario entender la estructura del negocio como lo establece en la primera fase del estándar CRISP-DM. A continuación veremos el modelo de datos que se sigue portafolio crediticio.

### 4.1.1 Modelo de Datos de la Empresa

Actualmente se cuenta con más de 35mil clientes con cartera activa en los 6 productos activos disponibles de los cuales están repartidos a más de 158

asesores de negocio en las 21 oficinas que cuenta la Caja Rural de Ahorro y Crédito los Andes.

La empresa cuenta un Data Center, donde se cuenta con servidores para el funcionamiento del Core Financiero Sifcnet de donde extraeremos todos los datos requeridos para las fases de compresión de los datos y preparación de los datos del estándar CRISP-DM.

### 4.1.2 Necesidades de Tecnologías de Información

Los sistemas de administración ofrecen cierta información que puede ayudar a la toma de decisiones en algunas ocasiones, sin embargo esta información es limitada y poco flexible. Dado que se busca el análisis de toda la información relevante de la empresa y el sistema de administración que maneja la empresa ofrece un análisis limitado, se requiere implementar las técnicas mencionadas en el capítulo II, para poder realizar un análisis con mayor profundidad que otorgue resultados con base en toda la información histórica de la empresa y no en análisis parciales de algunos segmentos de la información.

### 4.1.3 Recursos tecnológicos con los que cuenta

La Caja de Ahorro y Crédito Los Andes Los Andes cuenta con un servidor central para almacenar especialmente las aplicaciones desarrolladas llamado Andes Suite, por otro lado se cuenta con equipos de cómputo de las siguientes características para cada uno de los terminales que usaran el sistema (Gerentes, Analistas y Coordinadores):

- Sistema Operativo Windows 7.
- Procesador Core I5 de 3.4 Ghz.
- 4 gigas de memoria ram.
- Disco duro de 120 Gb.

### 4.1.4 Recursos tecnológicos que necesita

Se requiere una nueva terminal, con características similares a las que ya se tienen, que cuente con el sistema de inteligencia empresarial para realizar el análisis de la información o designar una de las terminales que no tenga tanta actividad operacional, para realizar esta tarea.

No sería conveniente designar al servidor también como terminal para el análisis de datos ya que se pueden entorpecer las actividades operacionales respecto al tiempo de respuesta y rapidez en las transacciones.

Una vez seleccionada el servidor en la que estará operando el sistema de inteligencia empresarial, se podrá implementar todo el proceso de inteligencia empresarial en los terminales requeridos.

### 4.1.5 Arquitectura del sistema

#### 4.1.5.1 Introducción

El caso de estudio es acompañado por el análisis y diseño del Sistema de Gestión de Portafolio Crediticio, identificado con el nombre de SGP y que sienta las bases para el desarrollo de trabajos futuros en el área. A continuación se presentan las diferentes secciones que definen la arquitectura inicial del sistema. Estas secciones tienen como guía el esquema de un documento de representación de arquitectura SAD [SAD].

#### 4.1.5.2 Propósito

Esta sección del informe tiene como propósito brindar una visión comprensible de la arquitectura general del SGP, utilizando diferentes vistas de la arquitectura para ilustrar diferentes aspectos del sistema. Captura las decisiones más importantes en lo que respecta a la arquitectura del sistema, obtenida durante su elaboración.

#### 4.1.5.3 Alcance

Se profundiza principalmente en las vistas de servicios y vista lógica, en donde se incluyen algunos elementos significativos al resto de las vistas. No se pretende dar una definición completa de la arquitectura sino una introducción que sirva como puntapié inicial a otras iteraciones dentro del estándar CRISP-DM.

#### 4.1.5.4 Representación de la Arquitectura

El sistema SGP está concebido como la base para dar respuesta a las interrogantes planteadas en el apartado uno del proyecto y que están relacionados con optimización de la gestión de portafolio crediticio.

La arquitectura está representada por diferentes vistas utilizando notación UML[UML03] de forma que se puedan visualizar, entender y razonar sobre los elementos de Gestión de Proyectos de Software significativos de la arquitectura e identificar las áreas de riesgo que puedan requerir mayor detalle de elaboración en propuestas futuras.

La siguiente sección detalla las vistas de la arquitectura que serán utilizadas para cubrir las dimensiones mencionadas, presentando a continuación el framework arquitectónico utilizado.

***Representación***

- Vista de Casos de Uso: Describe los procesos de negocio más significativos y el modelo del dominio. Presenta los actores y los casos de uso para el prototipo.

- Vista Lógica: Describe la arquitectura del prototipo presentando varios niveles de refinamiento. Indica los módulos lógicos principales, sus responsabilidades y dependencias.

- Vista de Deployment: Presenta aspectos físicos como topología, infraestructura informática.

***Framework Arquitectónico***

La arquitectura sigue el framework "4+1" (con variantes) presentado en [Kru95]; este framework define cuatro vistas para la arquitectura (4) en conjunto con los escenarios de uso (1), y es presentado en la siguiente figura:

***Figura N° 23**: Framework Arquitectónico*


**Fuente:** Metodología CRISP-DM

El mapeo de las vistas utilizadas a las propuestas en el framework se presenta en la siguiente tabla:

***Tabla N° 14:** Arquitectura en cada Framework*

| Framework 4+1 | Arquitectura |
|---|---|
| Use-Case View | Casos de Uso |
| Logical View | Lógica |
| Process View | Servicios |
| Implementación View | Datos |
| Deployment View | Deployment |

**Elaboración:** Por los investigadores.

#### 4.1.5.5 Vista de Casos de Uso

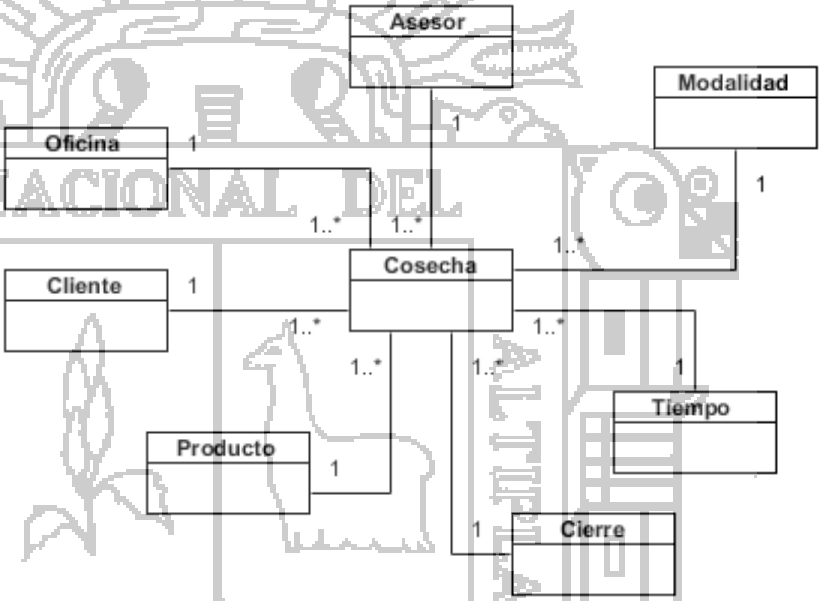
Esta vista presenta la percepción que tiene el usuario de las funcionalidades del sistema.

Se presentan los procesos de negocio más importantes, los casos de uso críticos que se derivan de éstos. Se muestran sus actores y se detallan los casos de uso significativos que tienen influencia sobre la arquitectura candidata.

**Modelo del Dominio**

El modelo dominio incluye el vocabulario necesario de comprender desde el punto de vista del problema y de la arquitectura.

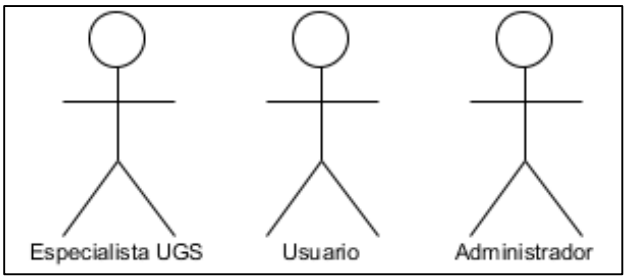
***Figura N° 24:*** *Modelo de Dominio*




**Actores**

Los actores son los que interactúan con el sistema y son los interesados en la problemática presentada en el informe.

***Figura N° 25****: Actores del Sistema.*




**Elaboración:** Por los investigadores.

Especialista UGS: Especialista en manejo del Sistema Sifcnet.

Usuario: Puede ser referente a los siguientes:

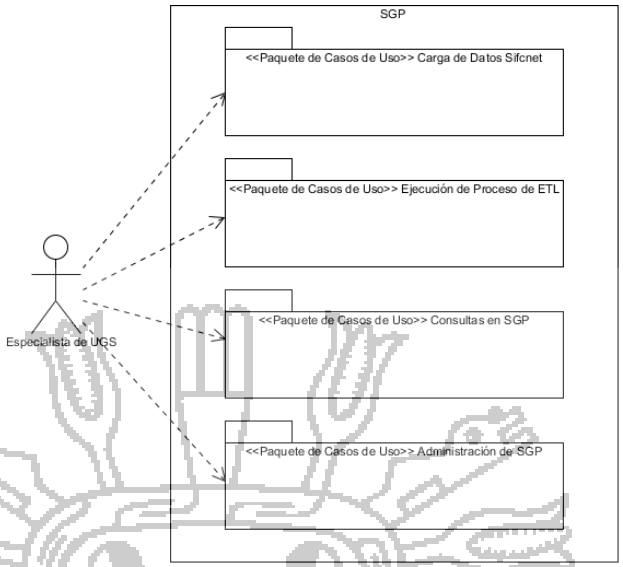
- Gerentes.
- Analistas.
- Coordinadores.

Administrador: Responsable de la administración de la solución sistema propuesta.

Los componentes y sistemas que forman parte de la arquitectura del prototipo son expresados en forma canónica.

El Modelo de Casos de Uso del Sistema se organiza en paquetes de casos de usos utilizados en el sistema SGP.

***Figura N° 26**: Modelo de Casos de uso del sistema*



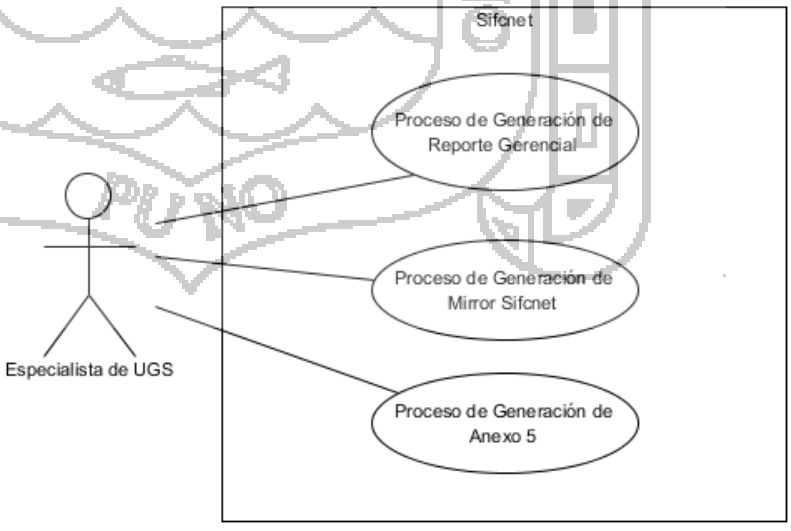

**Elaboración:** Por los investigadores.

Se atacan principalmente aquellos casos de uso relevantes para la lógica planteada en el caso de estudio.

**Casos de Uso**

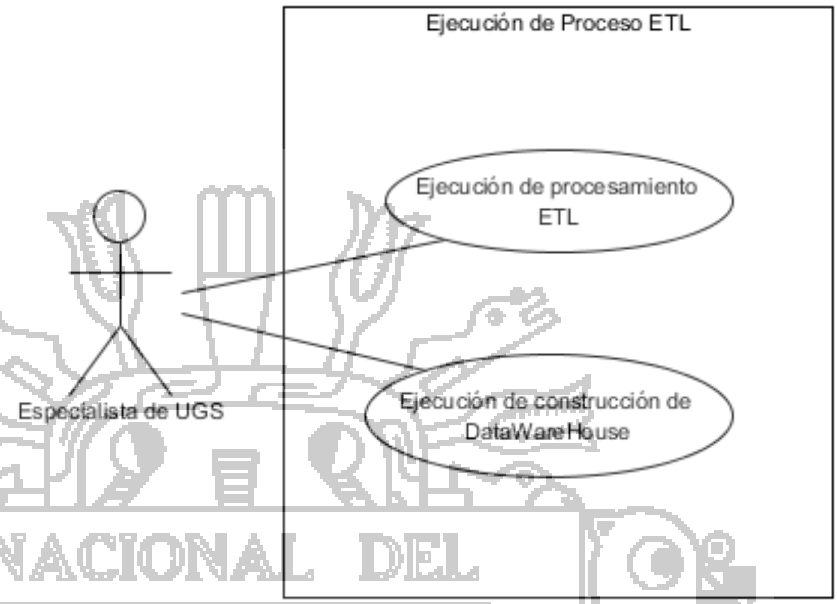
Los casos de uso del sistema son priorizados, se eligen los más importantes, y se deja para trabajos futuros completar el resto de los casos.

***Figura N° 27:** Diagrama de Caso de Uso, Carga de Datos Sifcnet*



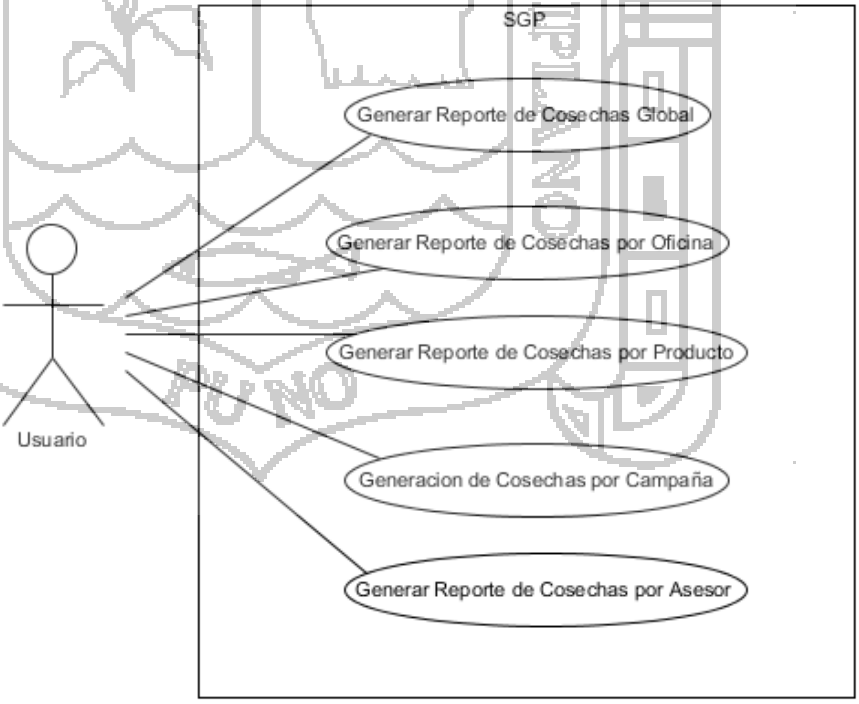

**Elaboración:** Por los investigadores.

***Figura N° 28:*** *Diagrama de casos de uso, Ejecución de Proceso ETL*



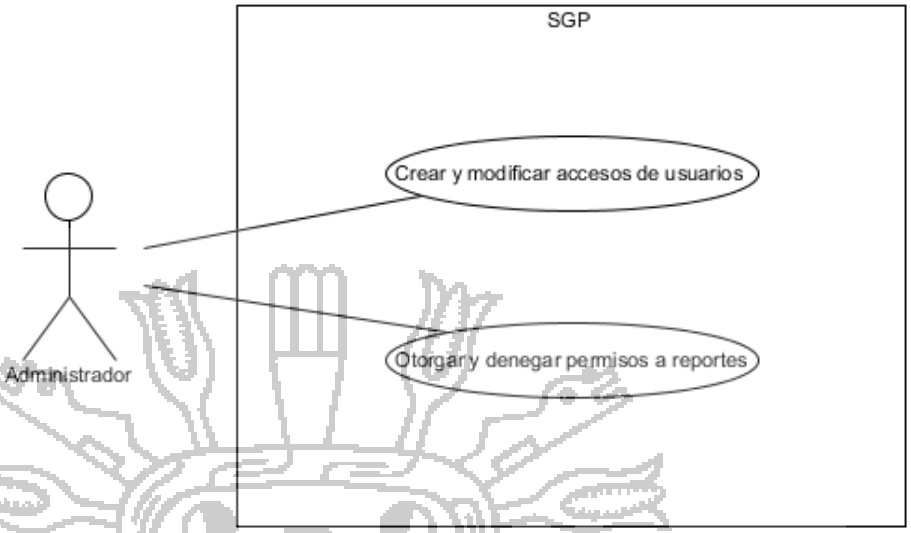

**Elaboración:** Por los investigadores.

***Figura N° 29****: Diagrama de casos de uso, consulta en el SGP*




**Elaboración:** Por los investigadores.

***Figura N° 30:*** *Diagrama de casos de uso, consultas al SGP nivel Administrador*




**Elaboración:** Por los investigadores.

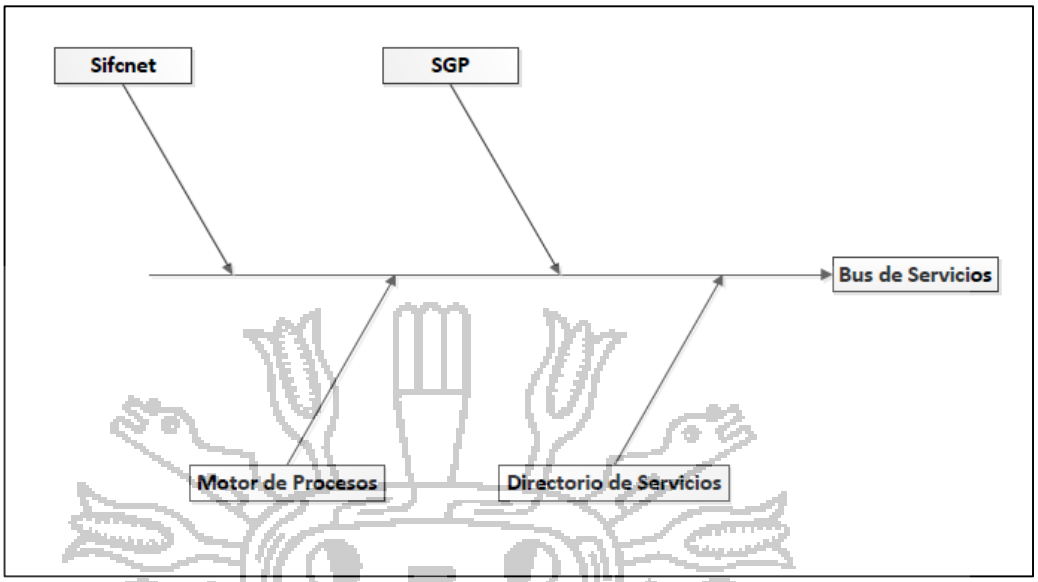
**Vista Lógica**

En este punto se presenta la vista lógica de la arquitectura a través de un refinamiento partiendo desde un nivel de abstracción mayor a uno menor. Como premisa la definición del prototipo está guiada por el patrón de arquitectura de orientación a servicios. Este patrón propone una definición de servicios independientes que ofrecen un conjunto de funcionalidades que son utilizadas por los diferentes actores de una forma altamente desacoplada, y con interfaces bien definidas.

**SGP**

En base a lo anterior, la representación de la arquitectura para el prototipo SGP utilizando el patrón de SOA, es el que se muestra en la figura.

***Figura N° 31:*** *Patrón de Arquitectura para el Sistema*




**Elaboración:** Por los investigadores.

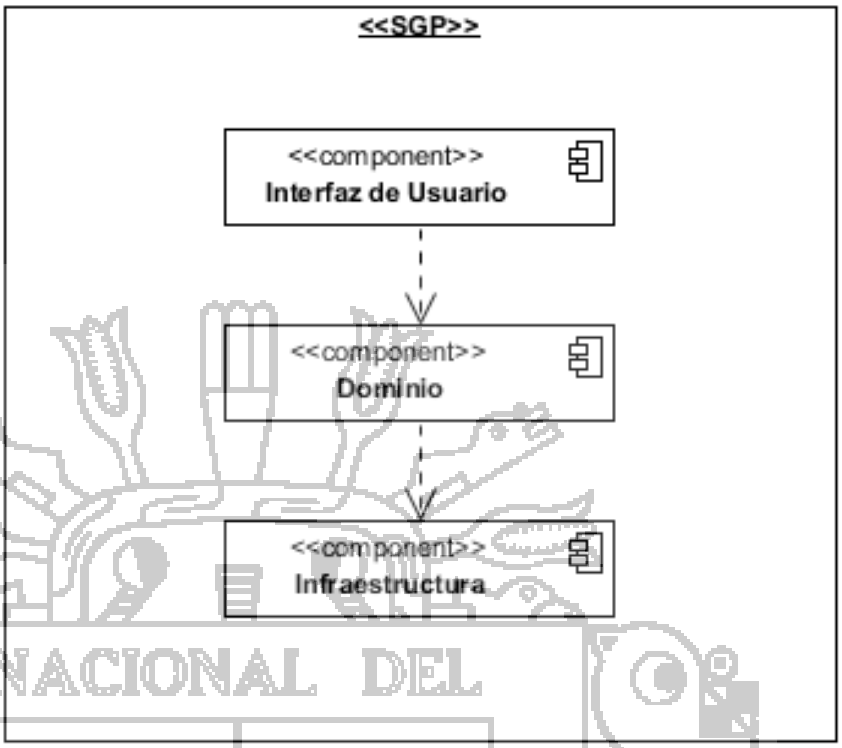
Todos los actores identificados participan de una arquitectura orientada a servicios, en donde se exponen a través del directorio de servicios cada uno de los servicios provistos.

El directorio de servicios es un proveedor de servicios que está encargado de almacenar los la información publicada por lo reportes generados del DWH los cuales se pueda consumir.

El bus de servicios es responsable de la comunicación entre los diferentes participantes del sistema, y de la administración de las interacciones entre los consumidores y proveedores, agregando funcionalidades de mayor nivel como puede ser la seguridad y mecanismo de transaccionalidad.

En la representación de la arquitectura se aprecia la inclusión de lo que se denomina motor de procesos, el cual en un contexto de reglas de negocios complejas para definir sus procesos asociados y facilitar la resolución de las reglas.

***Figura N° 32:*** *Arquitectura de SGP.*




**Elaboración:** Por los investigadores.

**Vista de Servicios**

En ésta vista de presentan los servicios tanto brindados como consumidos por los distintos actores del prototipo.

La siguiente tabla muestra los servicios identificados en el sistema, quien los provee y quien los consume:

***Tabla N° 15:*** *Servicios identificados para el sistema propuesto*

| | Provee | Consume |
|---|---|---|
| Sistema de Gestión de Portafolio | Conocimiento para la Toma de decisiones | Información Generada y Validada, Recursos tecnológicos. |
| Sifcnet | Registros | Información ingresada por interface. |

**Elaboración:** Por los investigadores.

**Vista de Datos**

En esta vista se presenta el modelo de datos utilizado en la elaboración del sistema.

**Modelo de Datos**

El modelo de datos presenta las relaciones existentes entre las principales entidades analizadas en el caso de estudio.

***Figura N° 33**: Modelo de Datos*

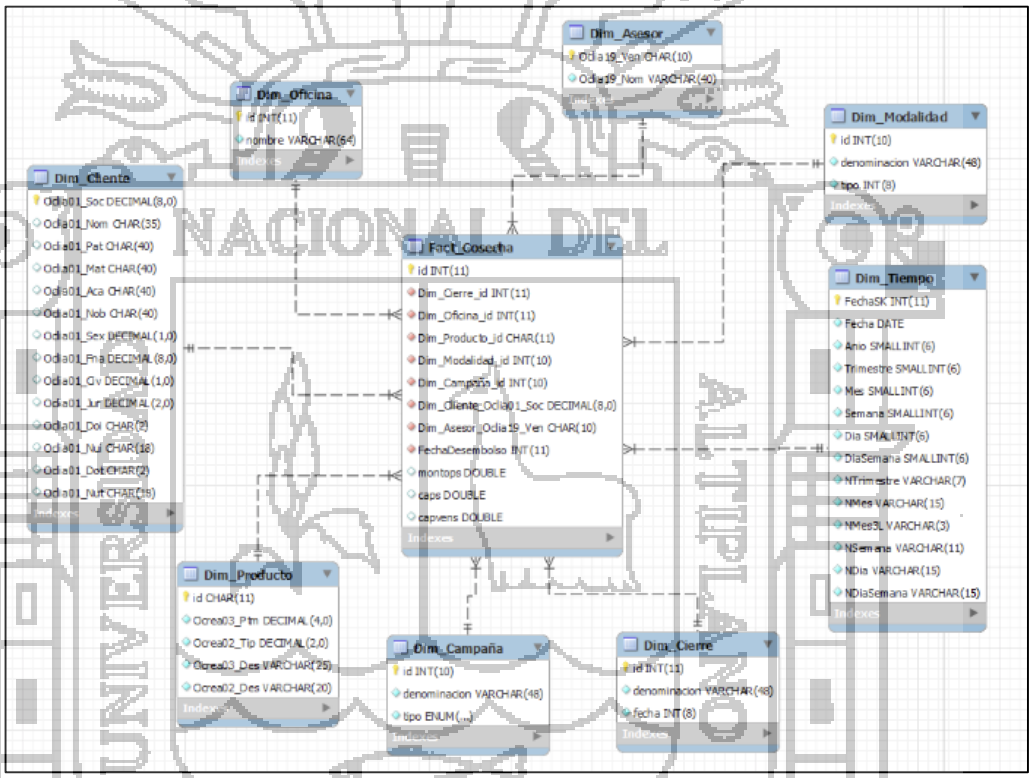



***Tabla N° 16:** Valores de los Atributos del Modelo multidimesional*

| | Atributo | Valor |
|---|---|---|
| Cliente | Oclia01_Soc | Decimal |
| | Oclia01_Nom | Char |
| | Oclia01_Pat | Char |
| | Oclia01_Mat | Char |
| | Oclia01_Aca | Char |
| | Oclia01_Nob | Char |
| | Oclia01_Sex | Decimal |
| | Oclia01_Fna | Decimal |
| | Oclia01_Civ | Decimal |
| | Oclia01_Jur | Decimal |
| | Oclia01_Doi | Char |

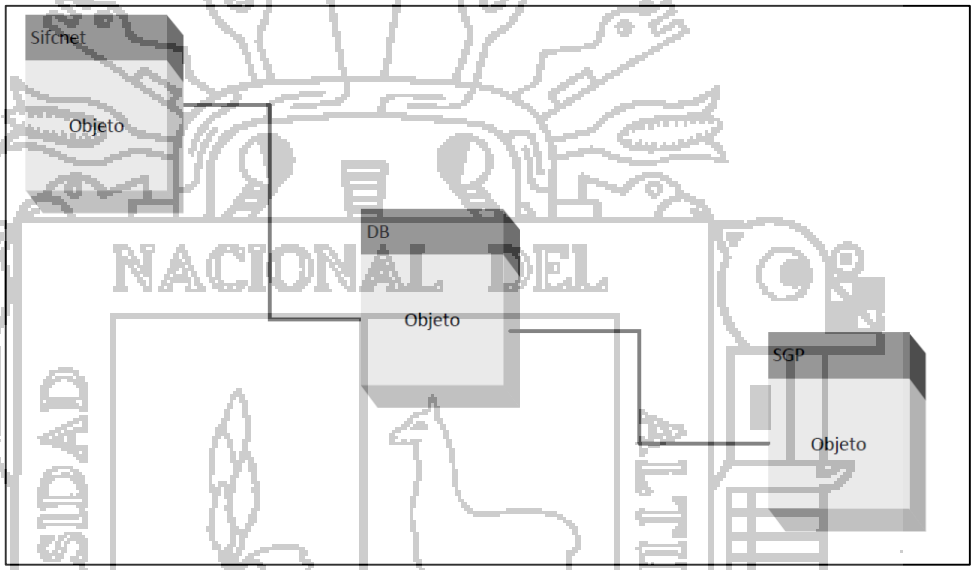
| | | |
|---|---|---|
| | Oclia01_Nui | Char |
| | Oclia01_Dot | Char |
| | Oclia01_Nut | Char |
| Oficina | Id | Int |
| | Nombre | varchar |
| Asesor | Oclia19_Ven | Char |
| | Oclia19_Nom | Varchar |
| Modalidad | Id | Int |
| | Denominacion | Varchar |
| | Tipo | Int |
| Producto | Id | Char |
| | Ocrea03_Ptm | Decimal |
| | Ocrea02_Tip | Decimal |
| | Ocrea03_Des | Varchar |
| | Ocrea02_Des | Varchar |
| Campaña | Id | Int |
| | Denominacion | Varchar |
| | Tipo | Enum |
| Cierre | Id | Int |
| | Denominacion | Varchar |
| | Fecha | Int |
| Tiempo | FechaSnk | Int |
| | Fecha | Decimal |
| | Año | SmallInt |
| | Trimestre | SmallInt |
| | Mes | SmallInt |
| | Semana | SmallInt |
| | Dia | SmallInt |
| | DiaSemana | SmallInt |
| | NTrimestre | Varchar |
| | NMes | Varchar |
| | NMes3 | Varchar |
| | NSemana | Varchar |
| | NDia | Varchar |
| | NDiaSemana | Varchar |
| Cosecha | Id | Int |
| | Cierre | Int |
| | Oficina | Int |
| | Producto | Char |
| | Modalidad | Int |
| | Campaña | Int |
| | Cliente | Decimal |
| | Asesor | Char |
| | FechaDesembolso | Int |
| | Montops | Double |
| | Caps | Double |
| | Capvens | Double |

**Elaboración:** Por los investigadores.

**Vista Deployment**

La vista de deployment presenta la infraestructura necesaria para soportar el sistema SGP. Se presenta aquí la arquitectura técnica de la aplicación indicando los nodos presentes en la infraestructura tecnológica esperada, y la localización de los componentes en dichos nodos.

***Figura N° 34:*** *Vista de Deployement.*




**Elaboración:** Por los investigadores.

La infraestructura cuenta con 3 nodos básicos. El primero (Sifcnet) cuenta con la información total relacionado con la estructura para la extracción previa de datos. Este puede tener varias instancias corriendo para la extracción de data y luego colocarla en (DB). Otro es el correspondiente al SGP, el cual está asociado a un único nodo, desde el cual se realizan todas las actividades de extracción de data ya procesada.

La infraestructura cuenta con 3 nodos básicos, el primero en donde se aloja el sistema Sifcnet desde donde se realiza la extracción de la información total de las oficinas, asesores, productos, créditos, etc, y el segundo nodo,

correspondiente al de DB Mirror y otros en donde reside la información ya procesada para múltiples fines que tiene la institución y por último el tercer nodo que aloja el DWH procesado y utilizado por SGP.

### 4.1.6 Modelo de datos multidimensional

En este caso no se mostrara el modelo multidimensional ya que es información confidencial de la entidad por lo cual solo expresaremos a nivel de descripción cada una de las tablas a utilizar en el proceso de construcción del dwh cosechas.

***Tabla N° 17:*** *Oclia01 - Clientes - Sifcnet*

| Alias | Formato | Longitud | Descripción |
|---|---|---|---|
| **Oclia01-Emp** | Numeric | 4 | Codigo de la empresa |
| **Oclia01-Soc** | Numeric | 8 | Codigo del cliente, incrementa a partir del ultimo registro de Oclia17 |
| **Oclia01-Emp0** | Numeric | 4 | Código de la empresa |
| **Oclia01-Nom** | Char | 35 | Nombre corto del cliente. Pat + Mat + Nob |
| **Oclia01-Pat** | Char | 40 | Apellido paterno (persona natural) ó Razón social (persona jurídica) |
| **Oclia01-Mat** | Char | 40 | Apellido materno (persona natural) ó Razón social (persona jurídica) |
| **Oclia01-Aca** | Char | 40 | Apellido de casada |
| **Oclia01-Nob** | Char | 40 | Nombres del cliente |
| **Oclia01-Dom** | Char | 40 | Domicilio del cliente |
| **Oclia01-Urb** | Char | 25 | Urbanización, barrio u otro tipo de grupo vecinal donde este ubicada la dirección |
| **Oclia01-Grp** | Char | 7 | Especificación de la ubicación de la dirección consignada |
| **Oclia01-Nur** | Char | 5 | Domicilio - Número |
| **Oclia01-Mza** | Char | 5 | Domicilio - Manzana |
| **Oclia01-Lot** | Char | 5 | Domicilio - Lote |
| **Oclia01-Int** | Char | 5 | Domicilio - Interior |
| **Oclia01-Emp1** | Numeric | 4 | Código de la empresa |
| **Oclia01-Dpt** | Numeric | 4 | Departamento donde se ubica el domicilio consignado |
| **Oclia01-Prov** | Numeric | 4 | Provincia donde se ubica el domicilio consignado |
| **Oclia01-Pos** | Numeric | 4 | Distrito donde se ubica el domicilio consignado |
| **Oclia01-Cas** | Numeric | 4 | |
| **Oclia01-Ing** | Numeric | 8 | Fecha de ingreso o registro del cliente en el SifcNet |
| **Oclia01-Fna** | Numeric | 8 | Fecha de nacimiento (Persona Natural) ó Fecha de inscripción a la SUNAT (Personas Jurídicas) consignada por el cliente |
| **Oclia01-Jur** | Numeric | 2 | Tipo de persona para la SBS |
| **Oclia01-Sex** | Numeric | 1 | Género o sexo del cliente (persona natural) |
| **Oclia01-Civ** | Numeric | 1 | Estado civil del cliente (persona natural) |

| | | | |
|---|---|---|---|
| **Oclia01-Ofi** | Numeric | 3 | Oficina donde se registro al cliente nuevo |
| **Oclia01-Rep** | Numeric | 8 | |
| **Oclia01-Emp2** | Numeric | 4 | Código de la empresa |
| **Oclia01-Loc** | Numeric | 3 | |
| **Oclia01-Tis** | Numeric | 2 | |
| **Oclia01-Ser** | Char | 20 | |
| **Oclia01-Act** | Numeric | 1 | |
| **Oclia01-Ret** | Numeric | 8 | Fecha de retiro del cliente |
| **Oclia01-Emp3** | Numeric | 4 | Código de la empresa |
| **Oclia01-Doi** | Char | 2 | Documento identidad del cliente que será registrado |
| **Oclia01-Nui** | Char | 18 | Número del documento de identidad (LE-Lib Electoral, CE-Carnet Extranj, FP-carnet policía, FA-car. F. Armadas, PA-Pasaporte) |
| **Oclia01-Emp4** | Numeric | 4 | Código de la empresa |
| **Oclia01-Dot** | Char | 2 | Documento tributario o Registro Unico del Contribuyente |
| **Oclia01-Nut** | Char | 18 | Número de Ruc |
| **Oclia01-Emp5** | Numeric | 4 | Código de la empresa |
| **Oclia01-Sec** | Char | 10 | Sectorista asesor del cliente (junto a la ruta son asignados posteriormente al cliente) |
| **Oclia01-Rut** | Char | 6 | Ruta del cliente (0 - NORMAL, 1 - PROBLEMA POTENCIAL, 2 - DEFICIENTE, 3 - DUDOSO, 4 - PERDIDA) |
| **Oclia01-Per** | Char | 5 | Tipo de persona para la SBS (01=PERSONA NATURAL, 02=PERSONA JURIDICA, 03=JUNTA SOLIDARIA, 04=CON. MANCOMUNADA, 05=ASOC. INDIVISA, 06=OTROS) |
| **Oclia01-Sep** | Char | 1 | Separador de bienes (Persona natural casada) S = Si, N = No |
| **Oclia01-Fec** | Numeric | 8 | Fecha en que se realizo la separación de bienes |
| **Oclia01-Pai** | Char | 4 | País origen o de nacimiento del cliente que será registrado (4006,4028,VE) |
| **Oclia01-Res** | Char | 1 | Reside en el pais de origen ( S = si, N = no ) |
| **Oclia01-Lab** | Char | 1 | Relación laboral con la entidad (T=trabajador, N=no tiene vínculo laboral, D=director, F=funcionario) |
| **Oclia01-Vin** | Char | 1 | Vínculo con la entidad (N=no vinculado, D=vinculación directa, I=Vinculación indirecta) |
| **Oclia01-Cio** | Char | 4 | Profesión u ocupación |
| **Oclia01-Gru** | Char | 5 | Grupo económico |
| **Oclia01-Te1** | Char | 10 | Teléfono 01 |
| **Oclia01-Te2** | Char | 10 | Teléfono 02 |
| **Oclia01-Te3** | Char | 10 | Celular 01 |
| **Oclia01-Te4** | Char | 10 | Celular 02 |
| **Oclia01-Fac** | Char | 40 | Facsímiles (Domicilio cliente) - dirección de negocios |
| **Oclia01-Dir** | Char | 1 | Tipo domicilio (1=Domicilio, 2=Domicilio Legal o Fiscal, 3=Taller o Negocio) |
| **Oclia01-Di1** | Char | 1 | |
| **Oclia01-Pro** | Char | 2 | Tipo de propiedad de domicilios |
| **Oclia01-Seq** | Numeric | 9 | |
| **Oclia01-Sbs** | Numeric | 9 | Código SBS |
| **Oclia01-Ciu** | Char | 7 | Actividad económica |
| **Oclia01-Of1** | Numeric | 2 | Registros Públicos: Sede |

| | | | |
|---|---|---|---|
| **Oclia01-Ced** | Numeric | 2 | Registros Públicos: Sub sede |
| **Oclia01-Tip** | Char | 1 | Registros públicos: Tipo registro (" " ó F = Ficha, P=partida, T=tomo) |
| **Oclia01-Num** | Char | 10 | Registros públicos: Número |
| **Oclia01-Res1** | Char | 1 | Residencia en la localidad (1=si es residente, 0=no es residente) |
| **Oclia01-Mag** | Numeric | 1 | Magnitud empresarial (1=Más de 10 000UIT, 2=600 a 10 000UIT, 3=300 a 600 UIT, 4=Menos de 300UIT) |
| **Oclia01-Acc** | Numeric | 1 | Accionista de la institución (1=si es accionista, 0=no es accionista) |
| **Oclia01-Dir2** | Numeric | 1 | |
| **Oclia01-Sig** | Char | 15 | Siglas Persona jurídica (Datos para entidades oficiales) |
| **Oclia01-Tot** | Decimal | 7 | Total de Hectáreas |
| **Oclia01-Cul** | Decimal | 7 | Área cultivable |
| **Oclia01-Uti** | Decimal | 7 | Área utilizable |
| **Oclia01-Mon** | Numeric | 3 | Moneda MONEDA NACIONAL = 0 |
| **Oclia01-Top** | Decimal | 15 | Monto de dinero |
| **Oclia01-Fel** | Numeric | 8 | Fecha |
| **Oclia01-Vtl** | Numeric | 8 | Fecha |
| **Oclia01-Sol** | Decimal | 15 | Monto de dinero |
| **Oclia01-Des** | Decimal | 15 | Monto de dinero |
| **Oclia01-Rem** | Decimal | 15 | Monto de dinero |
| **Oclia01-Fec2** | Numeric | 8 | Fecha de constitución(mancomunadas) |
| **Oclia01-Vto** | Numeric | 8 | Fecha |
| **Oclia01-Sit** | Numeric | 1 | Situación del cliente (bloqueado, vigente) |
| **Oclia01-Emp6** | Numeric | 4 | Código de la empresa |
| **Oclia01-Jun** | Numeric | 8 | Fecha |
| **Oclia01-Emp7** | Numeric | 4 | Código de la empresa |
| **Oclia01-Of2** | Numeric | 3 | Oficina |
| **Oclia01-Lib** | Numeric | 4 | |
| **Oclia01-Ti2** | Numeric | 1 | Tipo de persona |
| **Oclia01-Bco** | Numeric | 2 | VALOR POR DEFECTO '0' |
| **Oclia01-Ord** | Numeric | 5 | Números mayores a 10000 y 102412 campos con 0 |
| **Oclia01-Dig** | Numeric | 1 | 0-9 |
| **Oclia01-Cla** | Char | 4 | Espacios en blanco 38495, registros 0 = 63917 |
| **Oclia01-Con** | Numeric | 8 | Fecha |
| **Oclia01-Emi** | Numeric | 8 | Fecha |
| **Oclia01-Sit1** | Numeric | 1 | 0 y3 |
| **Oclia01-Fe1** | Numeric | 8 | VALOR POR DEFECTO '0' |
| **Oclia01-Mo1** | Char | 40 | |
| **Oclia01-Fe2** | Numeric | 8 | VALOR POR DEFECTO '0' |
| **Oclia01-Mo2** | Char | 40 | VALOR POR DEFECTO " (Vacío) |
| **Oclia01-Tra** | Numeric | 1 | 0, 1 y2 |
| **Oclia01-Cat** | Char | 1 | Garantías 0=persona jurídica, "espacio"=persona natural |
| **Oclia01-Bca** | Char | 1 | Garantías 0=persona jurídica, "espacio"=persona natural |
| **Oclia01-Cal** | Char | 1 | VALOR POR DEFECTO " (Vacío) |
| **Oclia01-Sca** | Char | 2 | VALOR POR DEFECTO " (Vacío) |

| | | | |
|---|---|---|---|
| **Oclia01-Hij** | Numeric | 3 | Número de hijos |
| **Oclia01-Rol** | Char | 6 | 0=persona jurídica, "espacio"=persona natural |
| **Oclia01-Mte** | Char | 1 | Envió Mensaje a Celular S=SI, N=NO |
| **Oclia01-Ope** | Decimal | 8 | Código del usuario del sistema |
| **Oclia01-Est** | Decimal | 8 | Estación |
| **Oclia01-Situ** | Numeric | 1 | Situación |
| **Oclia01-Fpr** | Decimal | 8 | Fecha de registro o modificación (yyyymmdd) |

**Elaboración:** Por los investigadores.

***Tabla N° 18:*** *Scona05 - Oficinas - Sifcnet*

| Alias | Formato | Longitud | Descripción |
|---|---|---|---|
| Cod_Empresa | decimal | 3, 0 | NOT NULL |
| Nom_Oficinal | char | 25 | NOT NULL |
| Nom_Oficinac | char | 15 | NOT NULL |

**Elaboración:** Por los investigadores.

***Tabla N° 19****: Oclia19 - Asesores – Sifcnet*

| Alias | Formato | Longitud | Descripción |
|---|---|---|---|
| Oclia19_Emp | decimal | 4 | NULL |
| Oclia19_Ven | char | 10 | NULL |
| Oclia19_Nom | char | 40 | NULL |
| Oclia19_Cco | decimal | 4,0 | NOT NULL |
| Oclia19_Sup | char | 10 | NULL |
| Oclia19_Usu | decimal | 8,0 | NOT NULL |
| Oclia19_Tel | char | 20 | NULL |
| Oclia19_Niv | char | 1 | NULL |
| Oclia19_Hab | char | 1 | NULL |
| Oclia19_Por | decimal | 6,3 | NOT NULL |
| Oclia19_Cap | char | 1 | NULL |
| Oclia19_Int | char | 1 | NULL |
| Oclia19_Cta | char | 30 | NULL |
| Oclia19_Mor | decimal | 6,3 | NOT NULL |
| Oclia19_Ctm | char | 30 | NULL |
| Oclia19_Tex | char | 1 | NULL |
| Oclia19_Pin | char | 8 | NULL |
| Oclia19_Ope | decimal | 10,0 | NOT NULL |
| Oclia19_Est | decimal | 10,0 | NOT NULL |
| Oclia19_Sit | decimal | 1,0 | NOT NULL |

| | | | |
|---|---|---|---|
| Oclia19_Fpr | decimal | 10,0 | NOT NULL |

**Elaboración:** Por los investigadores.

***Tabla N° 20:*** *Ocrea02 - Categoría Productos Activos - Sifcnet*

| Alias | Formato | Longitud | Descripción |
|---|---|---|---|
| Ocrea02_Emp | decimal | 4,0 | NOT NULL |
| Ocrea02_Tip | decimal | 2,0 | NOT NULL |
| Ocrea02_Des | char | 20 | NULL |
| Ocrea02_Ope | decimal | 6,0 | NOT NULL |
| Ocrea02_Est | decimal | 4,0 | NOT NULL |
| Ocrea02_Fpr | decimal | 8,0 | NOT NULL |

**Elaboración:** Por los investigadores.

***Tabla N° 21****: Ocrea03 - Productos Activos – Sifcnet*

| Alias | Formato | Longitud | Descripción |
|---|---|---|---|
| Ocrea03_Emp | decimal | 4,0 | NOT NULL |
| Ocrea03_Cla | decimal | 2,0 | NOT NULL |
| Ocrea03_Ptm | decimal | 4,0 | NOT NULL |
| Ocrea03_Des | char | 25 | NOT NULL |
| Ocrea03_Max | decimal | 15,2 | NOT NULL |
| Ocrea03_Min | decimal | 14,2 | NOT NULL |
| Ocrea03_Cmi | decimal | 4,0 | NOT NULL |
| Ocrea03_Cma | decimal | 4,0 | NOT NULL |
| Ocrea03_Int | decimal | 7,4 | NOT NULL |
| Ocrea03_In1 | decimal | 7,4 | NOT NULL |
| Ocrea03_In2 | decimal | 7,4 | NOT NULL |
| Ocrea03_Inn | decimal | 12,9 | NOT NULL |
| Ocrea03_Ti1 | decimal | 1,0 | NOT NULL |
| Ocrea03_Ti4 | char | 1 | NOT NULL |
| Ocrea03_Com | decimal | 7,4 | NOT NULL |

**Elaboración:** Por los investigadores.

### 4.1.7 Limpieza e integración de datos

Los datos almacenados en las bases de datos operacionales no siempre se encuentran homogéneos y estandarizados. Sobre todo cuando las base provienen de distintas fuentes. Esto se debe a que los datos hayan sido ingresados por diferentes personas, que no se haya definido con anterioridad un estándar para la captura de los datos o a simples errores humanos.

Para realizar un buen análisis tanto de OLAP como de data mining, es necesario que la información almacenada en el datawarehouse se encuentre lo más homogénea posible. Para esto se requiere pasar los datos por un proceso que permita la integración de los datos.

Para integrar la información de créditos de la Caja Rural de Ahorro y Credito Los Andes del caso de estudio tomaremos en cuenta las siguientes consideraciones:

Definición de tablas hecho, dimensión y sus respectivos atributos, que se tienen en la base de datos operacional, que serán usados para la implementación del datawarehouse y las transformaciones pertinentes para el transporte de los datos. Dicho nombres se establecieron como se muestra en la tabla 4.3.

**Tabla N° 22**: *Estándar de nombres definido para la implementación de datawarehouse*

| Elemento | Nombre | Ejemplo |
|---|---|---|
| Tablas | Nombre de la tabla en singular, primera letra de palabra con mayúscula luego todo minúsculas separados por "_" y sin acentos. | Oficina |
| Llaves | Nombre de la tabla en singular, primera letra en mayúscula luego minúsculas, guion bajo y nombre de atributo. | Dim_Cierre_id |
| Atributos | Primera letra en mayúscula las demás en minúscula, sin acentos. | Nombre |
| Dimensiones | Palabra "Dim" seguido de guion bajo y nombre de tabla. | Dim_Oficina |

| | | |
|---|---|---|
| Hechos | Palabra “Fact” seguido de guion bajo y nombre de tabla. | Fact_Cosecha |

**Elaboración:** Por los investigadores.

Verificar la inconsistencia que existen en los datos, ya que la base de datos de donde se extraerá la información carece de un modelo relacional por lo cual se pueden encontrar huecos en los registros en las tablas que contengan relación con otras tablas.

- Clientes que aparezcan en la tabla de Créditos pero no en la tabla de Clientes.
- Asesores de negocio que por cambios eventuales de cartera produzca incoherencias la tabla de cosechas.

### 4.1.8 Reportes necesarios

El reporte base extraído de la base de datos integral a nivel global de la institución reside en el RCD (Reporte Crediticio del Deudor), en donde se encuentra todo el registro base del deudor.

Este reporte es emitido cada fin de mes, donde se actualiza principalmente los saldos e intereses del deudor, a consecuencia de pagos de cuotas y otras operaciones.

- ¿Cuál es el Capital Vencido de los clientes que dejaron de pagar durante 30 días la cuota pactada durante identificados en cada cierre de mes?
- ¿Cuál es el monto desembolsado de los clientes identificados en cada cierre de mes?

### 4.1.9 Discusión final

Reside en cuanto se optimiza el tiempo y los procesos de la gestión de Portafolio Crediticio de los cuales también hemos definido todas las consideraciones que se tomaron en cuenta para la implementación del SGP en

cada una de las capas de actividades: ETL, OLAP y data mining. Se definió la arquitectura del sistema, el modelo de datos, los grupos de consultas de OLAP, la estructura de los archivos para data mining.

Todas estas actividades descritas a lo largo del capítulo corresponden a las primeras fases del estándar CRISP-DM: comprensión del negocio y compresión de los datos. Restan por cumplir 4 fases que veremos en el siguiente capítulo junto con la implementación del sistema y más detalles técnicos, así como la justificación de las herramientas seleccionadas.

## 4.2 Implementación

En este apartado se muestra la integración de las herramientas que forman al SGP, la justificación de su selección y detalles técnicos sobre la implementación del sistema. Se encuentra organizado de acuerdo a la arquitectura propuesta: en la sección 4.1.5.

Posteriormente los datos almacenados en el datawarehouse son explotados por la capa de análisis en donde se encuentran las técnicas de OLAP y data mining, por medio de las cuales se manipula la información para ofrecer resultados interesantes al usuario final sobre su propia información.

### 4.2.1 Creación de capa de integración

Para la creación de la capa de integración, se extrajeron los datos de la base de datos del Core financiero Sifcnet que se trabaja con el gestor de base de datos pervasive 10.0 y se almacenaran en el datawarehouse que se implementó en el manejador MySQL 5.0.

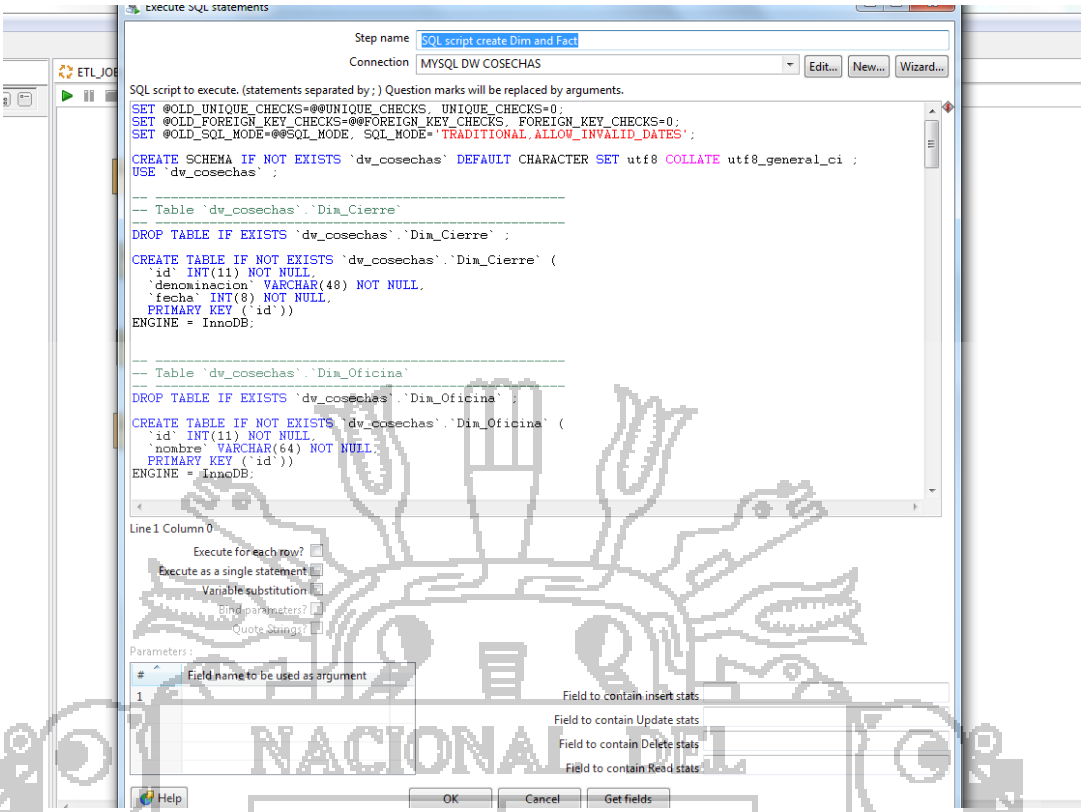
El gestor de base de datos Pervasive tiene una limitante de no ser un manejador de base de datos relacional, ya que carece de algunas características clásicas como el manejo de llaves foráneas y entidades referenciales.

Por otro lado, se decidió utilizar MySQL porque además de ser un manejador de código abierto, es uno de los más rápidos que existen para volúmenes de datos y por ello es ideal para ambientes de datawarehouse.

***Figura N° 35:*** *Proceso de Construcción de Estructura DW.*

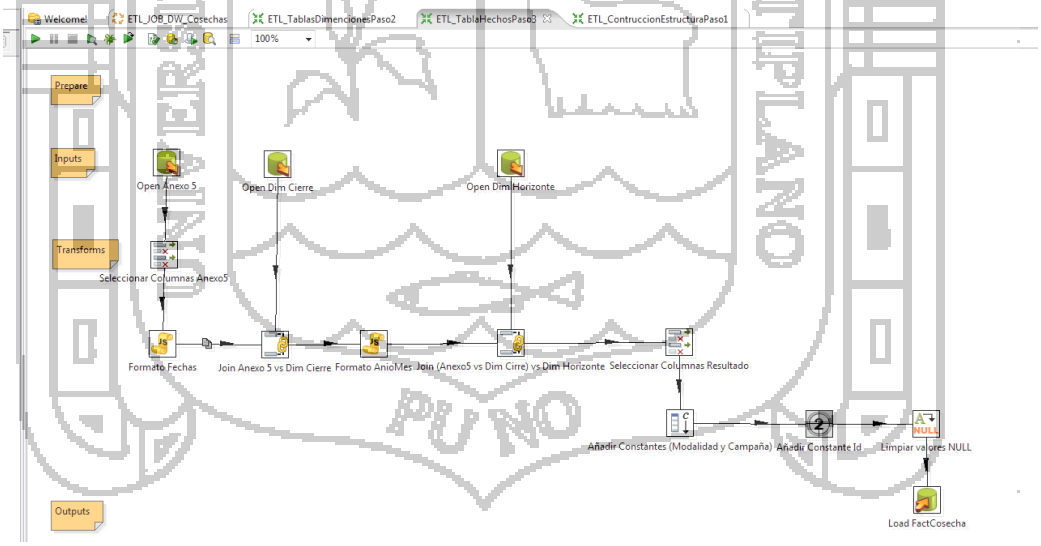

**Elaboración:** Por los investigadores.

***Figura N° 36:*** *Scripts SQL de Construcción de Estructura DW.*



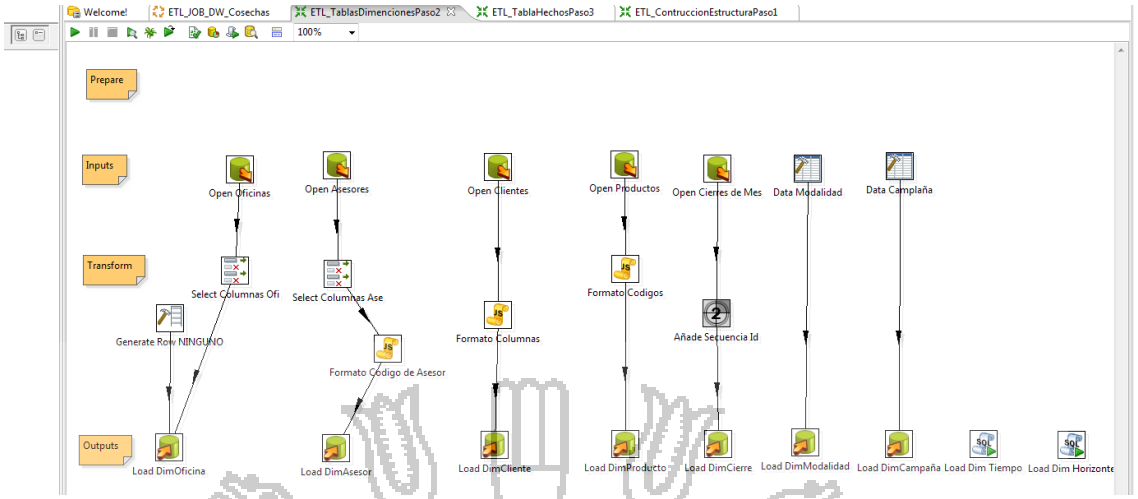

**Elaboración:** Por los investigadores.

***Figura N° 37:*** *Proceso de carga de Tabla de Hechos.*



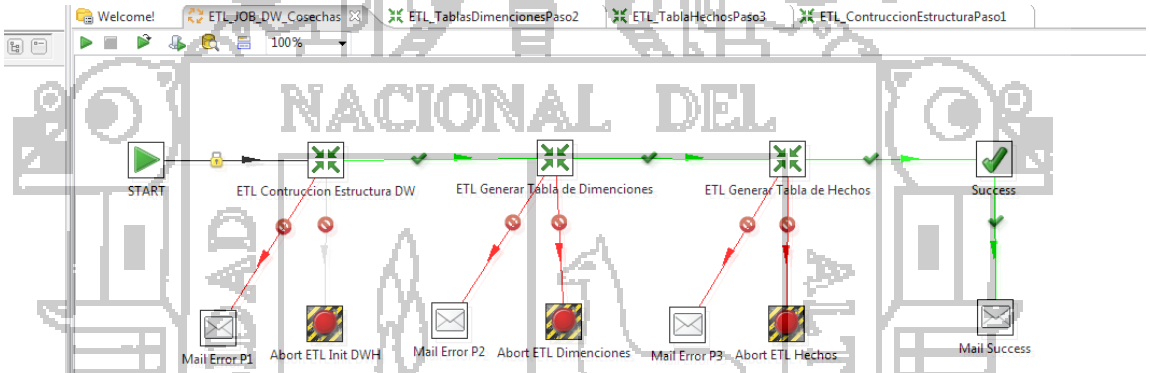

**Elaboración:** Por los investigadores.

***Figura N° 38****: Proceso de Carga de Tablas Dimensiones.*




**Elaboración:** Por los investigadores.

***Figura N° 39:*** *Programación del Proceso de Carga de DWH.*




**Elaboración:** Por los investigadores.

### 4.2.2 Creación de capa de análisis

#### 4.2.2.1 Mondrian

De las herramientas vistas en el capítulo II para aplicar OLAP a los datos almacenados en el datawarehouse se escogió Modrian.

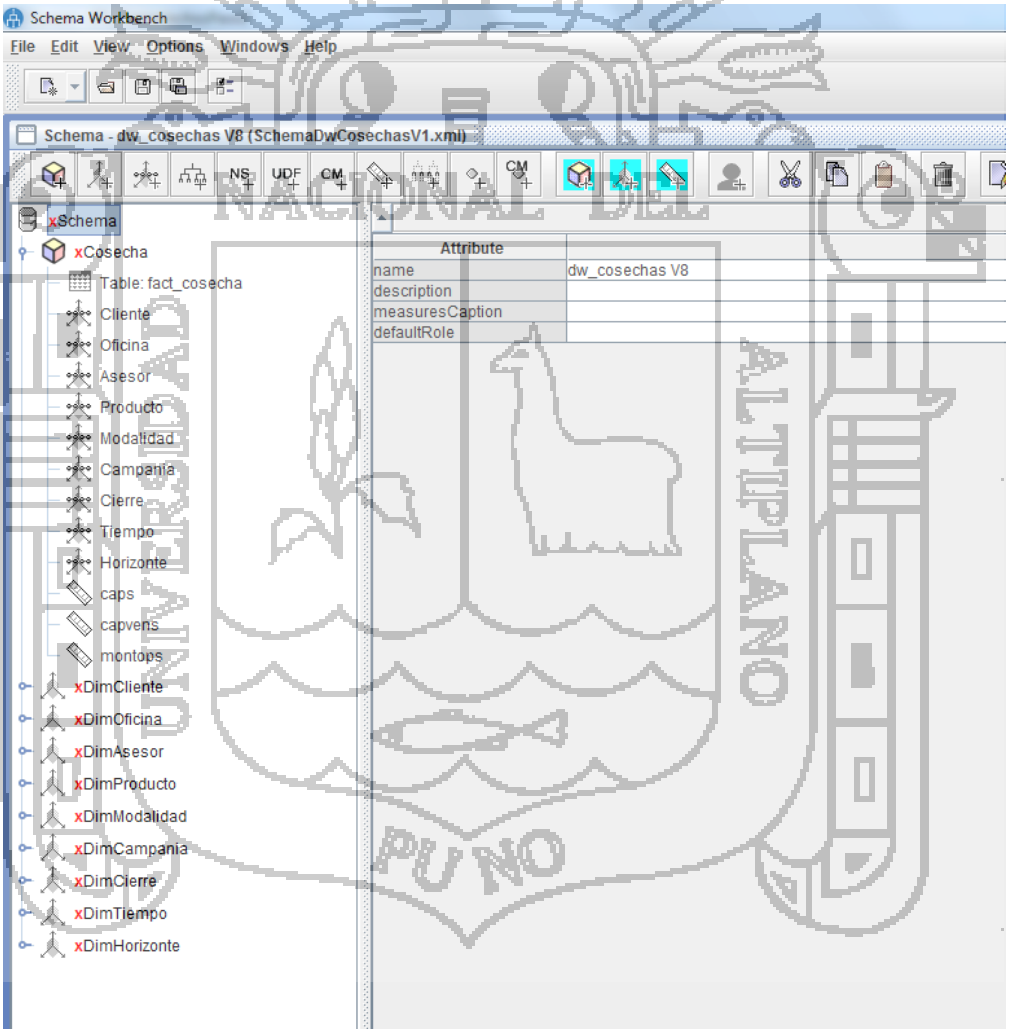
Las principales ventajas de Modrian provienen de su implementación ROLAP, ya que no tiene que generar cubos estáticos ahorrando el tiempo que cuesta generarlos y la memoria que ocupan.

Tiene la posibilidad de utilizar los datos que se encuentran en la base de datos, de esta manera se trabaja con los datos actualizados [Modrian, 2007]

Pese a que tradicionalmente los sistemas de OLAP implementados con MOLAP tienen una cierta ventaja de rendimiento, el uso de cache y de tablas agregadas que tiene Modria, hacen que se puedan obtener muy buenos rendimientos con él, sin perder las ventajas del modelo ROLAP.

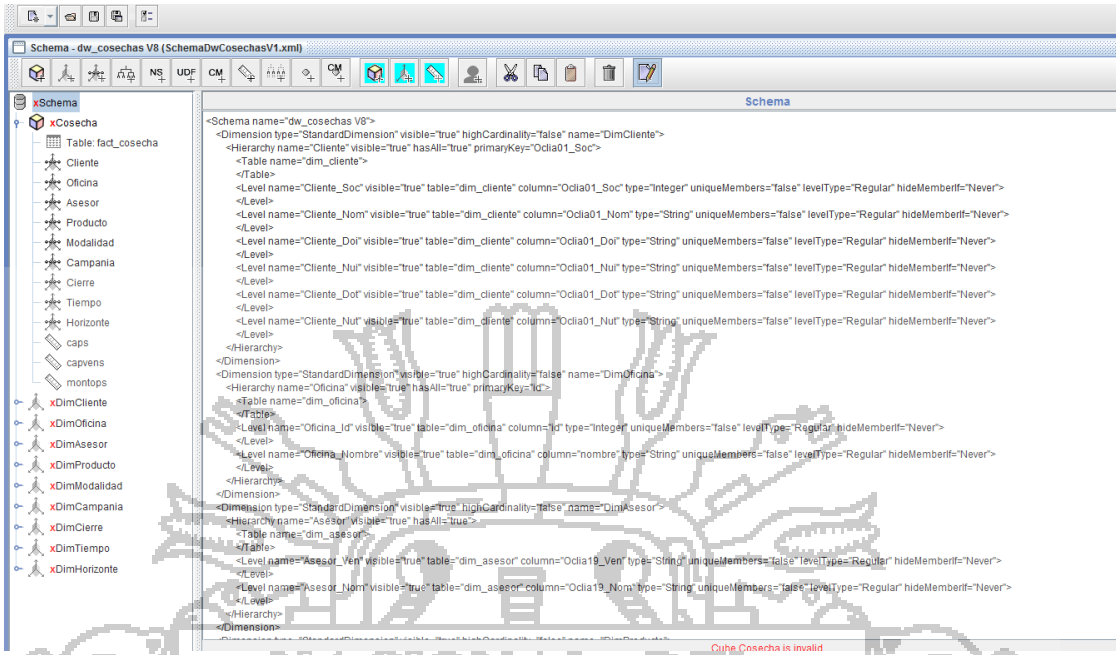
Para la elaboración del documento XML que representa al cubo de OLAP utilizaremos la herramienta Schema WorkBench.

***Figura N° 40:*** *SChema WorkBench Cubo OLAP Cosechas.*



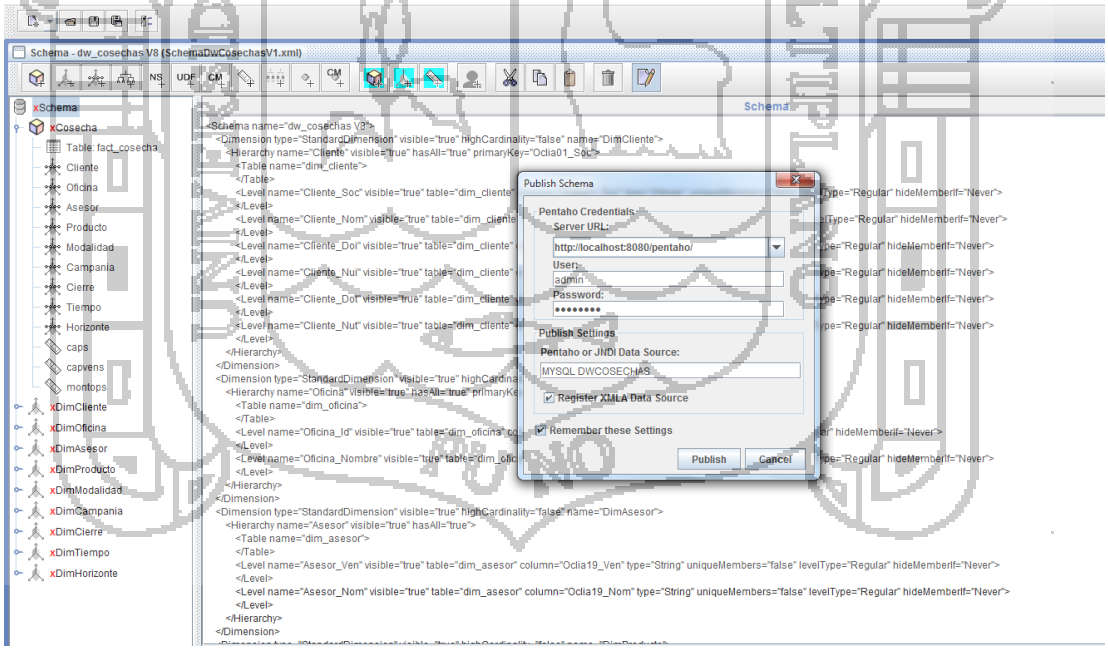

**Elaboración:** Por los investigadores.

***Figura N° 41: Schema WorkBench Xml que representa al cubo OLAP.***




**Elaboración:** Por los investigadores.

***Figura N° 42: Schema WorkBench Publicación de cubo OLAP a Pentaho Server-BI.***




**Elaboración:** Por los investigadores.

Ya que se a implementado el cubo de OLAP, podemos definir las consultas predeterminadas en el lenguaje MDX, las cuales responde a los grupos de consultas definidos en el capitulo IV. Se definieron 2 consultas principales que se muestran a continuación.

| Pregunta | Sentencia en lenguaje MDX |
|---|---|
| ¿Cuál es el Capital Vencido de los clientes que dejaron de pagar durante 30 días la cuota pactada durante identificados en cada cierre de mes? | WITH<br>SET [~FILTER] AS<br>{[Producto].[Producto_03Des].Members}<br>SET [~COLUMNS] AS<br>{[Cierre].[Cierre_Denominacion].Members}<br>SET [~ROWS] AS<br><br>Hierarchize({{[Horizonte].[Horizonte_AnioMes].Members}})<br>SELECT<br>NON EMPTY CrossJoin([~COLUMNS],<br>{{[Measures].[capvens]}}) ON COLUMNS,<br>NON EMPTY [~ROWS] ON ROWS<br>FROM [Cosecha]<br>WHERE [~FILTER] |
| ¿Cuál es el monto desembolsado de los clientes identificados en cada cierre de mes? | WITH<br>SET [~FILTER] AS<br>{[Producto].[Producto_03Des].Members}<br>SET [~COLUMNS] AS<br>{[Cierre].[Cierre_Denominacion].Members}<br>SET [~ROWS] AS<br><br>Hierarchize({{[Horizonte].[Horizonte_AnioMes].Members}})<br>SELECT<br>NON EMPTY CrossJoin([~COLUMNS],<br>{{[Measures].[montops]}}) ON COLUMNS,<br>NON EMPTY [~ROWS] ON ROWS<br>FROM [Cosecha]<br>WHERE [~FILTER] |

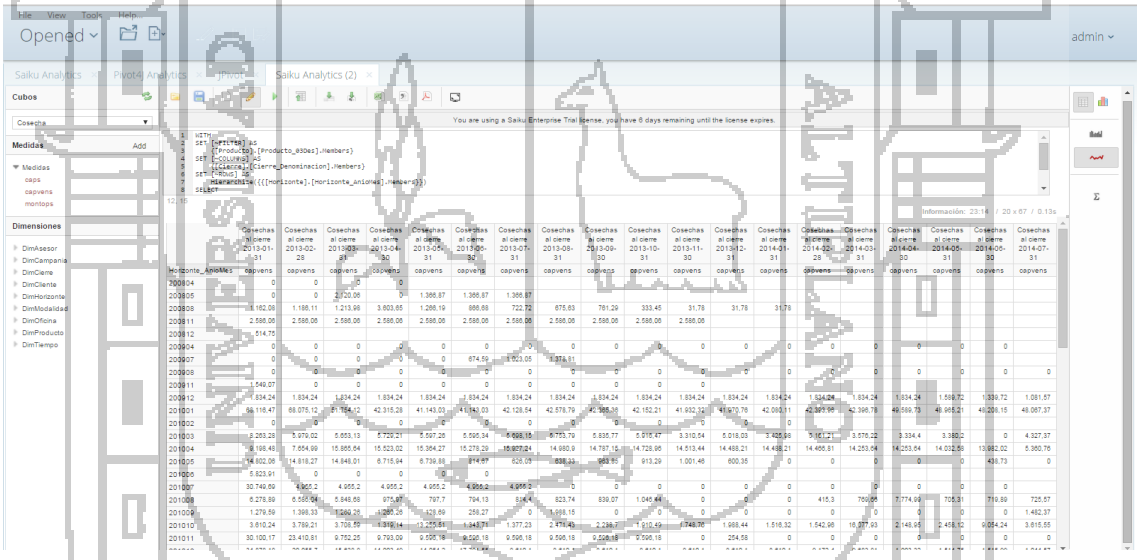
**Elaboración:** Por los investigadores.

Es importante recordar que las consultas en MDX son mapeadas posteriormente a lenguaje SQL, para obtener datos requeridos desde el datawarehouse que se encuentra implementado en un manejador de base de datos realcional.

A continuacion se muestra un manejador de analisis de datos Saiku Analytics, antes de ello debemos loguearnos.

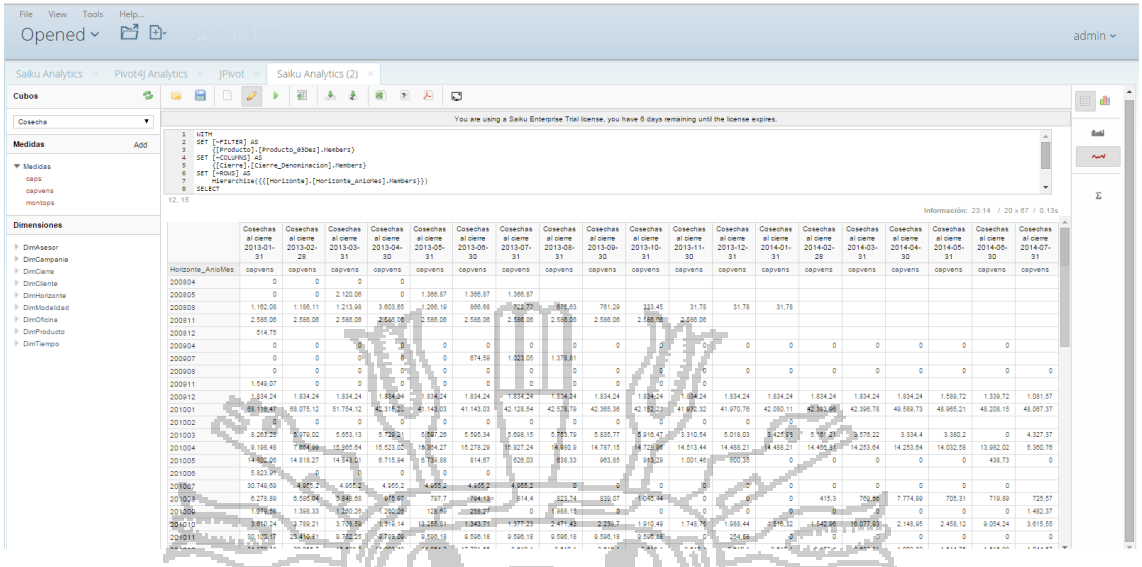
***Figura N° 43: Inicio de Sesión en Pentaho Server-BI.***


***Figura N° 44:** Análisis de Datos con Saiku Analytics.*



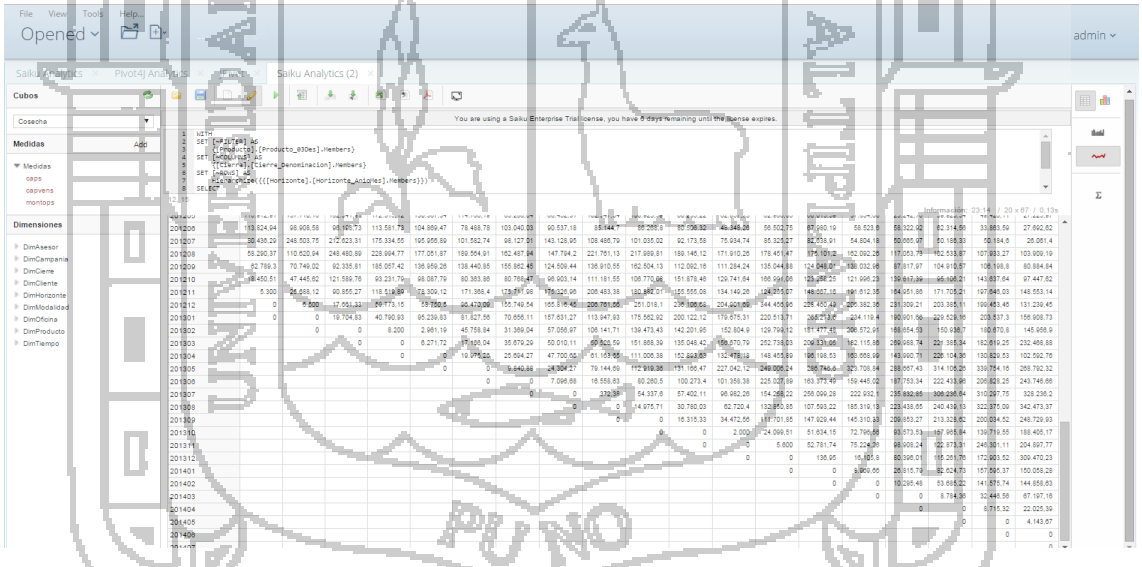

**Fuente:** Pentaho.

***Figura N° 45:*** *Análisis de Datos con Saiku Analytics.*



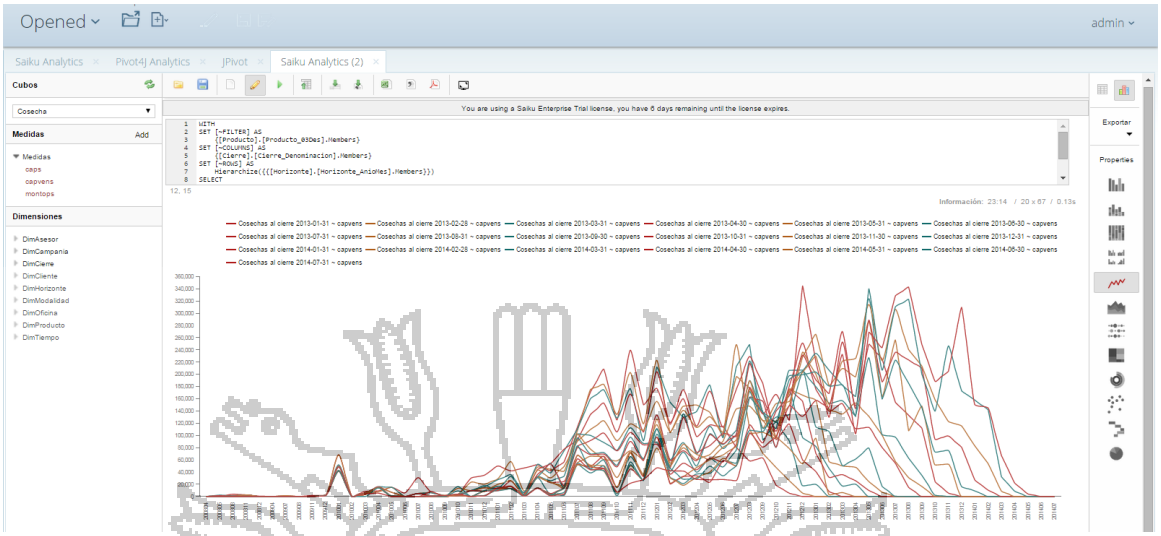
**Fuente:** Pentaho.

***Figura N° 46:*** *Análisis Datos con SaIKU Modelamiento de los Datos.*



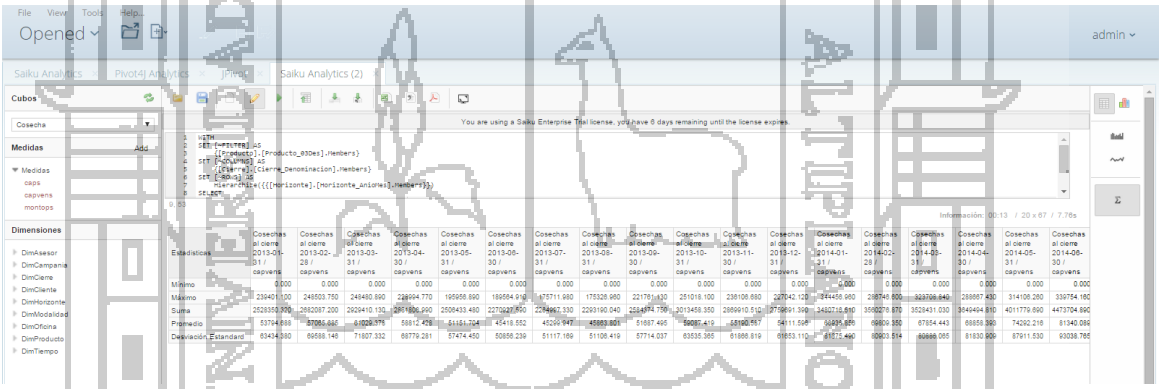
**Fuente:** Pentaho.

***Figura N° 47**: Análisis Datos con Saiku analytics Generación de conocimiento.*



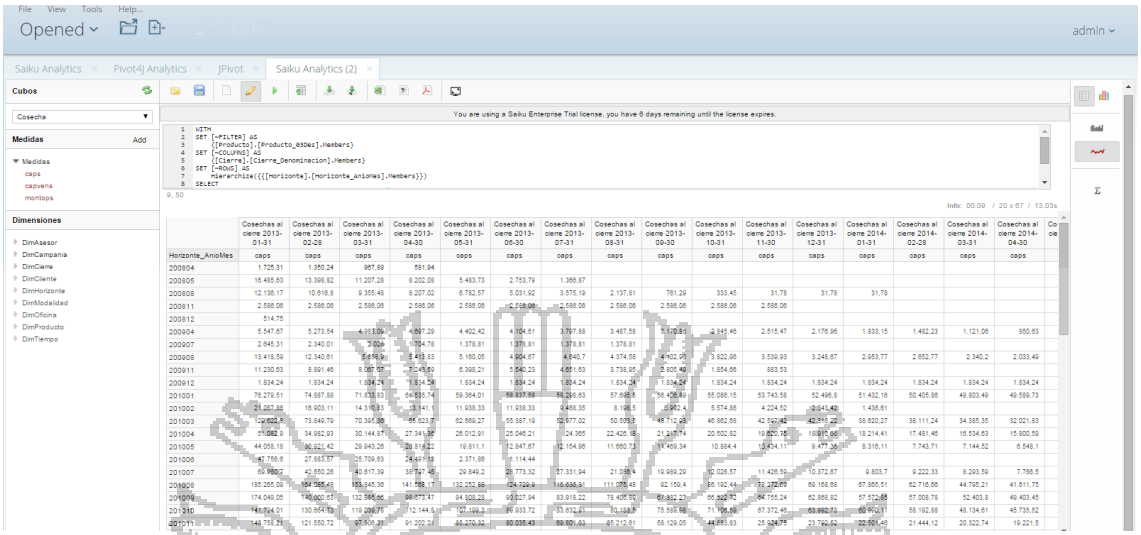

**Fuente:** Pentaho.

***Figura N° 48**: Análisis Datos con Saiku Analytics.*



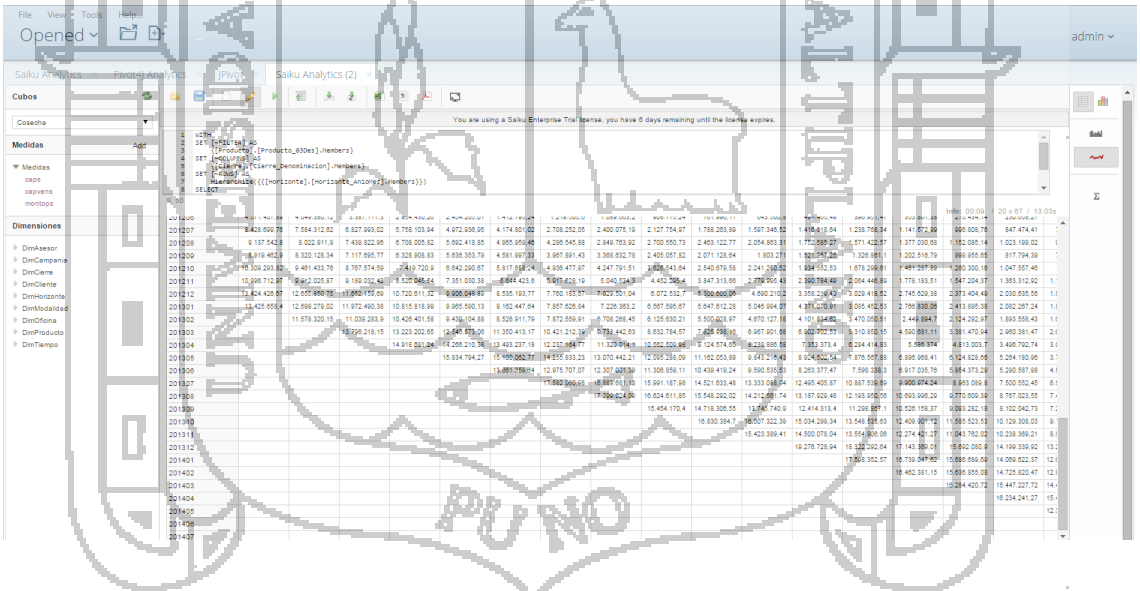

**Fuente:** Pentaho

***Figura N° 49**: Análisis Datos con Saiku Analytics.*



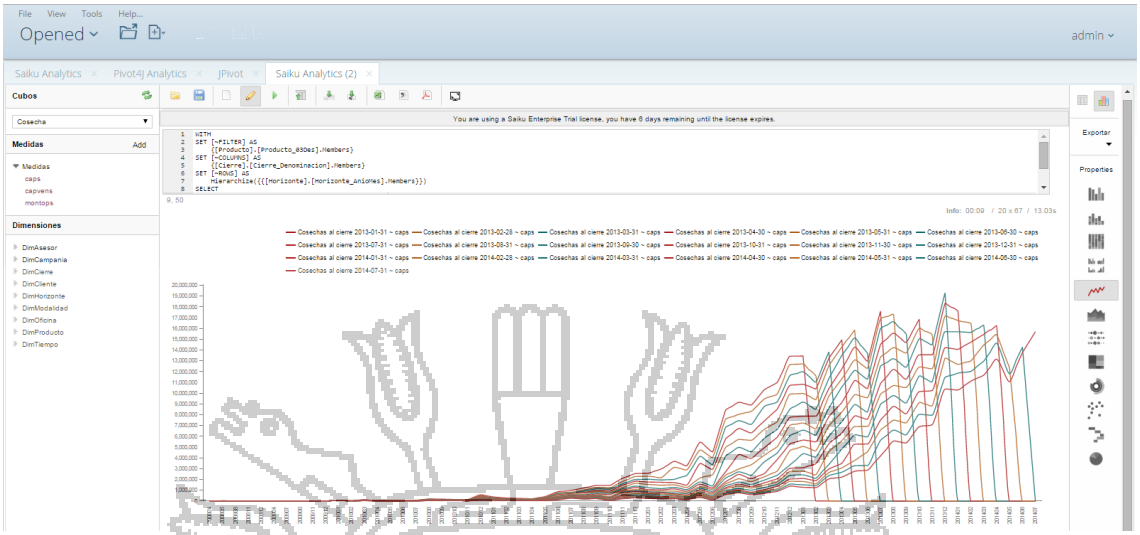
**Fuente:** Pentaho.

***Figura N° 50**: Análisis Datos con Saiku Generación de Conocimiento y modelamiento de Reportes.*



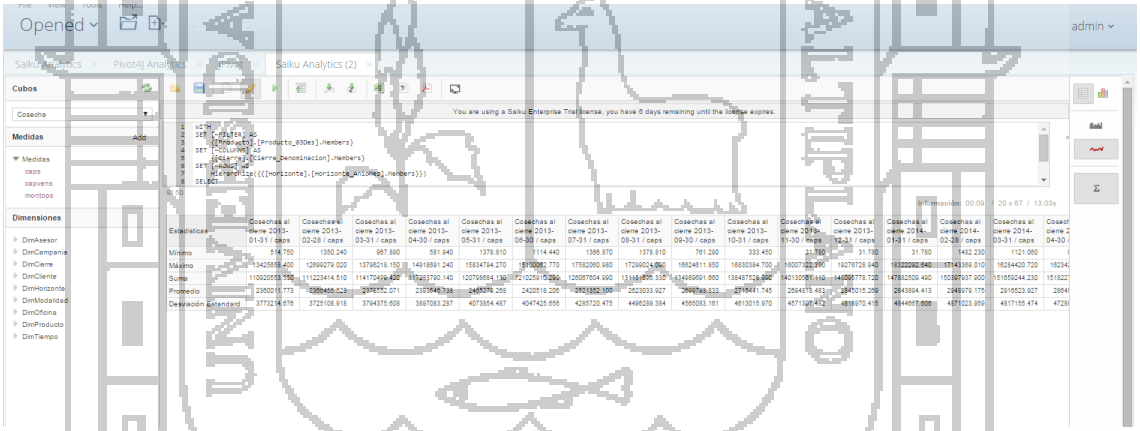
**Fuente:** Pentaho.

***Figura N° 51**: Análisis Datos con Saiku Analytics.*




**Fuente:** Pentaho.

***Figura N° 52**: Análisis Datos con Saiku Analytics.*




**Fuente:** Pentaho.

## 4.3 Optimización del Pre procesamiento de datos.

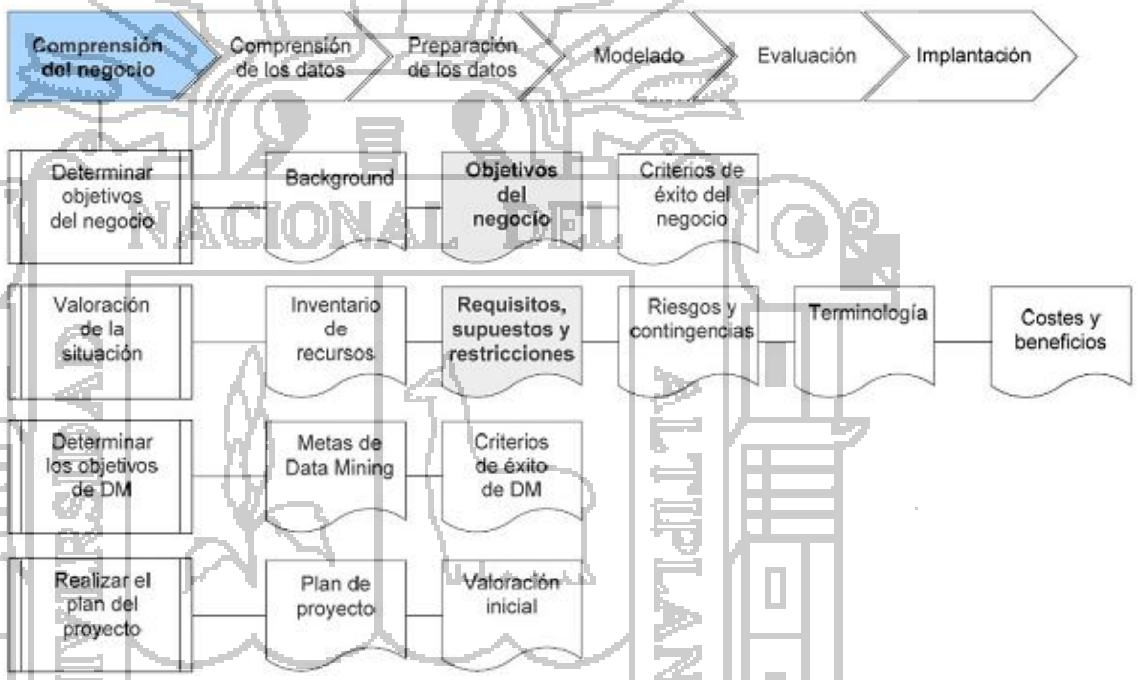
### 4.3.1 Requerimientos para la implementación

Bajo la notación BPM se determinó el proceso manual que se realiza para la Gestión de Portafolio, y se establece los requerimientos, los cuales son:

- Una Base de Registros o Base de Datos actuales a un punto actual del tiempo (mes de cierre).

- Una Preparación de los registros para establecer un historial de los mismos registros, ordenados por el tiempo.
- Presentación de agrupaciones de registros bajo la dimensión de tiempo.
- Creación de reportes para el entendimiento de usuarios.
- Análisis de los reportes.

Estos requerimientos son contrastados con la Metodología optada; CRISP-DM.




**Fuente:** Metodología CRISP-DM

Y se observa similitud en los procesos, y se determina que lo requerimientos para el proceso rudimentario actual, pueden ser llevados o traducidos en la metodología presentada.

### 4.3.2 Optimización de Preparación de Datos o ETL.

La metodología para la prueba, dada en el punto 3, es la del contraste de la pre prueba y la post prueba.

Se enumeran las pruebas con la muestra ya establecida:

**Pre Prueba N° 1.**

| Proceso | Tiempo |
|---|---|
| Requerimiento de Información | 2 Horas. |
| Preparación de Información | 18 Horas. |

**Pre Prueba N° 2.**

| Proceso | Tiempo |
|---|---|
| Requerimiento de Información | 3 Horas. |
| Preparación de Información | 17 Horas. |

**Pre Prueba N° 3.**

| Proceso | Tiempo |
|---|---|
| Requerimiento de Información | 2.5 Horas. |
| Preparación de Información | 19 Horas. |

**Post Prueba N° 4.**

| Proceso | Tiempo |
|---|---|
| Requerimiento de Información | Implícito |
| Preparación de Información | 1.0 Horas. |

**Post Prueba N° 5.**

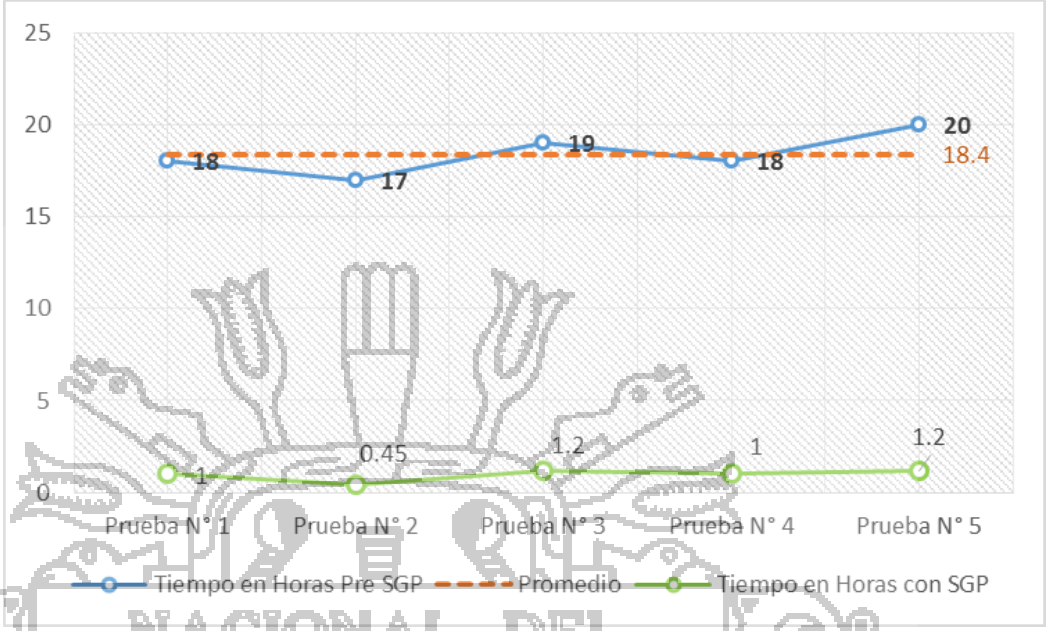
| Proceso | Tiempo |
|---|---|
| Requerimiento de Información | Implícito. |
| Preparación de Información | 1.2 Horas. |

**Resultados de la Pre Prueba.**

| Proceso | Tiempo Promedio |
|---|---|
| Requerimiento de Información | Implícito |
| Preparación de Información | 0.97 Horas. |

Los cuadros anteriores son las pruebas utilizando el sistema. Aplicando la observación de las pruebas sin el sistema, y con el sistema denotamos las mejoras

***Grafico N° 4:*** *Resultados en la pruebas con SGP, en el proceso ETL inmerso.*




**Elaboración:** Por los investigadores.

Es evidente la mejora aplicando las pruebas con el Sistema de Portafolio propuesto, tal como se observa en el Grafico N° 4. Se mejora de 18.4 a 1.17 en promedio de horas.

## CONCLUSIONES

El informe comprende en una primera instancia el estudio del estado de la implementación de automatización de procesos que cumplan con la Gestión de Portafolio de Créditos, y específicamente en el Sub proceso de este, que es la Generación de Cosechas Crediticias. Los cuales siguen procesos rudimentarios, nada propio de una gestión corporativa de información masiva.

Se plantea Optimizar El proceso de Gestión de Portafolio utilizando la tecnología Data Mining, y se llega a las siguientes conclusiones.

**PRIMERO:** Dado el Caso de estudio se establecen los requerimientos tales como información necesaria dado en los campos de las base datos para los hechos y dimensiones, que forman parte del cubo para la minería de datos. Al Igual que el Sistema de Gestión de Portafolio Requiere de Software especializado integral como es el que se dio uso.

**SEGUNDO:** Se lograron optimizar 5 procesos de la gestión de portafolio crediticio, logrando disminuir de 18.4 hora promedio a 0.97 horas en pre procesamiento de datos.

**TERCERO**: En el caso de estudio se detalla que la metodología CRISP-DM permite la implementación del Sistema de Gestión de Portafolio Crediticio, por ser modelado bajo la metodología que plantea.

**CUARTO**: Se demostró que el Sistema de Gestión de Portafolio Crediticio logra generar conocimiento para el análisis y toma de decisiones.

## RECOMENDACIONES

**PRIMERO**: Se recomienda seguir con este tipo de soluciones para los siguientes componentes de la Gestión de Portafolio Crediticio, que no se contempló en esta investigación.

**SEGUNDO**: Se recomienda utilizar herramientas de software distintas a Windows ya que podemos ver mayores ventajas en cuanto a costos.

**TERCERO**: Se recomienda mejorar el acceso de usuarios y la masificación del uso del Sistema de Gestión de Portafolio Crediticio.

**CUARTO**: Se recomienda diseñar un interface con mejor interactuación y que requiera menos especialización por parte del usuario.

## BIBLIOGRAFÍA

AVILA, A. R. (2001).Metod*ología de la investigación*. Lima – Perú: Estudios y ediciones RA.

BERSON A, & SMITH S. (1997); *Data Warehousing, Datamining and OLAP*: Me Graw Hill.

CASILLAS, R. C. (2004). *Desarrollo de Aplicaciones Web.* Barcelona: Eureca.

CASILLAS, R. C. (2005). *Bases de Datos.* Barcelona: Eureca.

CÓRDOVA, Z. M. (2003).*Estadística descriptiva e inferencial*. Lima – Perú: Moshera.

HERNANDEZ, S.; FERNÁNDEZ, C. & BAPTISTA, L. (2003).*Metodología de la investigación*. Mexico: McGrawHill.

HERNANDEZ, J.; RAMIREZ, J.; FERRI, C. (2004). Introduccion a la Mineria de Datos. Lima - Perú: Pearson Prentice Hall.

HUMPHRIES M. & HAWKINS W. (1999). *Data Warehousing: Architecture and Implementation*: Prentice Hall.

JACOBSON I, B. G. (2000). *El Proceso Unificado de Desarrollo de Software.* Madrid: Pearson Education.

JIAWEIL H. & KAMBER M. (2001). *Data Mining Concepts and Techniques*. Morgan Kaufmann Publisher.

KENDALL, K. &. (2005). *Análisis y Diseño de Sistemas.* México: Pearson Education.

LARMAN, C. (2003). *UML y Patrones.* Madrid: Prentice Hall.

MUKESH M. (1999). *Data Warehousing and Knowledge Discovery*: Ed. Springer.

PRESMAN, R. (2002). *Ingeniería del Software: un enfoque práctico.* Madrid: McGraw- Hill.

SOMERVILLE, I. (2005). *Ingeniería del Software.* Madrid: Pearson Education.

UOC. (2004). *Desarrollo de Aplicaciones Web.* España: Eureca.

SALTON M. (1993). *Introduction to Modern Information Retrieval*: McGraw Hill.

Super Intendencia de Banca y Seguros. (2014). Portal de Gobierno. Recuperado 17 de Junio 2014 de. http://www.sbs.gob.pe/.

Inkae Negocios. (2012). Minería de Datos. Recuperado 18 Junio 2014 de http://inkanegocios.com/viajes/?p=772.

GALLARDO A. J. (2007). *Metodología para el Desarrollo de Proyectos en Minería de Datos CRISP-DM.*

RUMBAUGH J, & JACBSON I, & BOOCH G. (2000). *El Lenguaje Unificado de Modelado, Manual de Referencia:* **Addison Wesley.**

GARIMELLA K, & LEES M, & WILLIAMS B. *BPM, Gerencia de Procesos de Negocio.* Recuperado 17 de Diciembre 2014 de http://www.konradlorenz.edu.co/images/publicaciones/suma_digital_sistemas/bpm.pdf.

**ANEXOS**

Anexo 1: Cuestionario de la Pre –Prueba.

Anexo 2: Matriz de Consistencia.

**Anexo 1**: Encuesta de Pre – Prueba.

1.- ¿Como considera Ud. La gestion actual de portafolio de creditos?

Muy Bueno ☐ Bueno ☐ Regular ☐ Deficiente ☐ Muy Deficiente ☐

2.- ¿Cómo considera el tiempo de entrega de resultados de la gestion actual de portafolio?

Muy Bueno ☐ Bueno ☐ Regular ☐ Deficiente ☐ Muy Deficiente ☐

3.- ¿Cómo considera Ud. La vizualiacion de los reportes entregables de la gestion?

Muy Bueno ☐ Bueno ☐ Regular ☐ Deficiente ☐ Muy Deficiente ☐

4.- ¿Cómo considera Ud. La gestion de portafolio con la herramienta Excel?

Muy Bueno ☐ Bueno ☐ Regular ☐ Deficiente ☐ Muy Deficiente ☐

5.- ¿Según apreciacion, es adecuado la gestion de portafolio con hojas de calculo?

Muy Bueno ☐ Bueno ☐ Regular ☐ Deficiente ☐ Muy Deficiente ☐

6.- ¿Considera Ud. Que deberia implementarse un Sistema de Gestion de Portafolio Crediticio?

Muy Bueno ☐ Bueno ☐ Regular ☐ Deficiente ☐ Muy Deficiente ☐

**Anexo 2:** Matriz de Consistencia

**Proyecto de Investigación Experimental, Sistemas de Información.**

**Título:** Optimización del Proceso de Gestión de Portafolio Crediticio con la Implementación de un Sistema de Gestión con el uso de Data Mining.

| Planteamiento del Problema | Objetivos | Hipótesis | Variables e Indicadores | Muestra | Diseño. |
|---|---|---|---|---|---|
| **Pregunta General** | **Objetivo General** | **Hipótesis General** | **Variable Independiente** | **Población** | **Método** |
| ¿Qué efectos produce el Sistema de Gestión de Portafolio Crediticio con el uso de Data Mining en la gestión de portafolio de la Caja Rural de Ahorro y Crédito Los Andes? | Optimizar el proceso de gestión de portafolio crediticio con la implementación de un sistema de gestión utilizando la tecnología Data Mining para la elaboración de Cosechas Crediticias que derive en toma de decisiones en la Caja Rural de Ahorro y Crédito Los Andes. | El sistema de gestión utilizando tecnología Data Mining optimiza produce efectos positivos en la gestión de portafolio crediticio para la toma de decisiones en la Caja Rural de Ahorro y Crédito Los Andes. | *Sistema de Gestión de Portafolio Crediticio.* | La población está constituida por los registro de datos ingresados desde el año 2008 hasta el mes de Junio del 2014. | ***Método:*** Experimental ***Área de Investigación:*** Sistemas de Información<br><br>$O_1$: Encuesta de opinión sobre proceso actual de gestión de portafolio crediticio.<br><br>X: Sistema de Gestión utilizando tecnología Data Mining.<br><br>$O_2$: Encuesta de opinión sobre proceso de gestión de portafolio |
| **Pregunta Especifica** | **Objetivos Específicos** | **Hipótesis Especificas** | **Variable Dependiente** | **Muestra** | |
| a) ¿Cuáles son los requerimientos | a) Determinar los requerimientos para el desarrollo | a) Se obtendrá los requerimientos de ejecutivos de la | - Proceso de Gestión de Portafolio Crediticio. | La muestra está constituido por 3 analistas de riesgos y 6 | |

| | | | | | |
|---|---|---|---|---|---|
| que se deben determinar para el desarrollo del Sistema de Gestión de Portafolio Crediticio?<br>b) ¿Qué debemos identificar para optimizar el Pre procesamiento de datos para la Gestión de Portafolio Crediticio, en la elaboración de las Cosechas crediticias.<br>c) ¿Cómo Determinar el Modelo de negocio acorde a los | del Sistema de Gestión de Portafolio Crediticio.<br>b) Identificar y optimizar el Pre procesamiento de datos para la Gestión de Portafolio Crediticio, en la elaboración de las Cosechas crediticias.<br>c) Determinar el Modelo de negocio acorde a los procesos de Gestión de Portafolio en la elaboración de las cosechas.<br>d) Generar Conocimiento como resultado del proceso de gestión | Caja Rural de Ahorro y Crédito Los Andes para el crecimiento corporativo, que viabilice el desarrollo del sistema de gestión de portafolio crediticio en la elaboración de cosechas crediticias.<br>b) El proceso de preparación de datos realizados manualmente serán optimizados por el componente de Data Mining- ETL.<br>c) La metodología de desarrollo CRISP-DM permite el modelamiento de | | Gerentes. Esta selección fue de tipo no probabilístico.<br>Es decir n=8, donde n: Muestra | crediticio utilizando el sistema de gestión utilizando tecnología Data Mining. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | procesos de Gestión de Portafolio en la elaboración de las cosechas?.<br>d) ¿Cómo se debe Generar Conocimiento como resultado del proceso de gestión aplicando Data Mining? | aplicando Data Mining. | negocio para la implementación del Sistema de Gestión de Portafolio Crediticio.<br>El Sistema de Gestión de Portafolio Crediticio como resultado generara conocimiento para la toma de decisiones. | | | |