**II CONGRESO INTERNACIONAL DE PROCESAMIENTO DE LA INFORMACIÓN**

**Modelo de replicación de datos para sistemas de información distribuidos**

***Data replication model for distributed information systems***

**Gloria Raquel Leyva Jerez1, Osmel Pérez Alzola2, Nayibi Martín Peña3**

1-Gloria Raquel Leyva Jerez. Universidad de las Ciencias Informáticas, Cuba. E-mail: grleyva@uci.cu

2-Osmel Pérez Alzola. Universidad de las Ciencias Informáticas, Cuba. E-mail: opereza@uci.cu

3-Nayibi Martín Peña. Universidad de las Ciencias Informáticas, Cuba. E-mail: nmartin@uci.cu

**Resumen:** La descentralización de la información en los diversos sectores de la sociedad ha propiciado ambientes distribuidos, aparejado al uso de herramientas de replicación. La replicación de datos es más que una simple copia de datos entre diferentes almacenamientos; comprende el análisis, diseño, implementación, administración y monitorización de un servicio que debe garantizar la consistencia de los datos replicados. Precisamente, la presente investigación propone un modelo de réplica de datos que contribuye a mantener la consistencia de los datos replicados en ambientes distribuidos, proceso sustancial en sistemas de información tales como: hospitalarios, gubernamentales, bancarios, entre otros como empresariales, con alto flujo de datos. La propuesta comprende principios, características, componentes que describen procesos vinculados con la replicación de datos, modelados con la técnica estratégica de Gestión de Procesos de Negocio (BPM) y que pueden implementarse con tecnologías modernas y libres; y por indicaciones metodológicas para su aplicación. La propuesta fue validada con métodos científicos, cuantitativos y cualitativos, entre ellos: Likert, Iadov, Grupos focales, Triangulación metodológica y método de prueba. Las validaciones realizadas evidencian que existe una correspondencia satisfactoria entre el objetivo y los resultados obtenidos. La aplicación de la propuesta en el desarrollo del Replicador de datos REKO corroboró la contribución del modelo para mantener la consistencia de los datos, y constituye un aporte a la soberanía tecnológica y a la sustitución de importaciones.

***Abstract:*** *Decentralization of information in the various sectors of society has led to distributed environments, coupled with the use of replication tools. Data Replication is more than a simple copy of data between different storages. It includes the analysis, design, implementation, administration and monitoring of a service that must guarantee the consistency of the replicated data. Precisely, this research proposes a data replication model that contributes to maintaining the consistency of replicated data in distributed environments; important process in information systems such as: hospital, government, banking systems, among others as enterprises, with a high data flow. The proposal includes principles, characteristics, components that describe processes related to data replication, modeled with the strategic technique of Business Process Management (BPM) and that can be implemented with modern and free technologies; and by methodological indications for its application. The proposal was validated with quantitative and qualitative methods, among them: Likert, Iadov, Focus groups, Methodological triangulation and test method; evidencing that there is a satisfactory correspondence between the objective and the results obtained. The validations carried out show that there is a satisfactory correspondence between the objective and the results obtained. The application of the proposal in the development of the REKO Data Replicator corroborated the contribution of the model to maintain data consistency, and it is a contribution to technological sovereignty and import substitution.*

**Palabras Clave:** Bases de datos distribuidas; Consistencia; Gestión de procesos de negocio; Réplica de datos; Sistemas de información.

***Keywords:*** *Distributed databases; Consistency; Business process management; Data replication; Information systems.*

**1. Introducción**

La consistencia de los datos tiene un carácter decisivo en cualquier organización, ello se puede ilustrar con sencillos ejemplos que representan situaciones comunes [1, 2]:

- Administración de medicamentos en un hospital donde cualquier modificación en los registros electrónicos del tratamiento de un paciente podría tener gravísimas consecuencias para su salud.

- En la gestión de incentivos desde el Departamento de Recursos Humanos donde modificar o eliminar datos dentro de la información de nóminas puede conllevar a un impago o a practicar un cómputo inexacto.

- En la elaboración de un reporte gerencial donde si se carece de la integridad necesaria entre sistemas, esta falta de integridad afectará a sus relaciones, en primer lugar, lo que provocará que el reporte no pueda beneficiarse de la inteligencia de negocio y no sea, en ningún caso, fidedigno ni válido.

Como se ha ejemplificado la consistencia es un aspecto clave en la replicación [3, 4]. Si los sistemas de información no emplean una solución de réplica que garantice la consistencia de la información se desencadenan problemas de integridad, afectando la toma de decisiones y la calidad de los procesos gestionados en los sistemas de información. No tomar las decisiones correctas puede conducir a problemas en el rendimiento, el funcionamiento, la calidad de los servicios ofrecidos y la elevación de los costos de desarrollo y mantenimiento.

Como parte fundamental del marco teórico de la investigación se realizó un diagnóstico sobre el conocimiento explícito abordado en la siguiente literatura especializada, tanto académica con los modelos homólogos examinados, como empresarial mediante las soluciones comerciales de replicación de datos: Modelo para la experimentación en técnicas de actualización de réplicas [5], Modelo de replicación de datos en sistemas distribuidos de bases de datos relacionales [6], Modelo de servicios web para replicación de instantáneas sobre motores múltiples de bases de datos [7], Modelo de replicación de Microsoft SQL Server [8], Oracle Streams [9], pglogical [10], SymmetricsDS [11], y Tungsten Replicator [12]. Las soluciones se analizaron en base a un grupo de criterios (procesos de réplica, aplicabilidad, software libre, evaluación de la consistencia) identificados en la literatura y ratificados por expertos en el tema a través de una encuesta aplicada a especialistas de la UCI con experiencia en el desarrollo y despliegue de soluciones de replicación de datos, pertenecientes a los centros: Consultoría y Desarrollo de Arquitecturas Empresariales (CDAE), Tecnologías de Gestión de Datos (DATEC), Dirección de Informatización (DIN) y Tecnologías para la Formación (FORTES).

El estudio realizado mediante una análisis documental y comparativo entre modelos y soluciones de replicación de datos evidenció un conjunto de insuficiencias que atentan contra la capacidad de los modelos para contribuir a mantener la consistencia en la replicación de datos, debido a que:

- La documentación de calidad que se encuentra publicada en la mayoría de los casos pertenece a compañías líderes que se refieren específicamente a las características de sus herramientas. Algunas soluciones guían a los equipos en “qué” hacer sobre determinados procesos, pero no especifican “cómo” hacerlo. Esto provoca falta de entendimiento en los analistas y desarrolladores, que limita las capacidades que contribuyen directamente a lograr los resultados esperados.

- La variedad de marcos de trabajo y tecnologías para el desarrollo de soluciones de réplica dificulta la toma de decisiones en el proceso de selección y para sugerir un método que gestione la consistencia de los datos.

- Ninguna de las soluciones y modelos describe en su la totalidad los procesos lo cual limita las capacidades de los sistemas de replicación para responder de manera eficaz ante cada proceso.

- Todos los casos no evidencian explícitamente cómo gestionan la consistencia en la distribución de los datos, en consecuencia, las capacidades de las soluciones para ofrecer un servicio de calidad con alto grado de consistencia se ven limitadas.

Dada la situación antes expuesta, se define el siguiente **problema de la investigación** a resolver: ¿Cómo contribuir a mantener la consistencia de los datos replicados en ambientes distribuidos? A partir del problema enunciado, se plantea como **objetivo general de la investigación**: Desarrollar un modelo para la réplica de datos en sistemas de bases de datos distribuidas que contribuya a mantener la consistencia de los datos replicados.

**2. Metodología**

La vida del modelo está definida por la dinámica de su funcionamiento y su aplicabilidad, por lo que se fue perfeccionando su construcción y desarrollo a medida que se elaboraba.

**2.1 Principios y características del modelo propuesto**

Para superar las insuficiencias anteriormente detectadas y con el propósito de desarrollar un modelo de replicación de datos con capacidad para contribuir a mantener la consistencia en la distribución de los datos en ambientes distribuidos, fueron definidos los siguientes principios:

- Orientado a contribuir a mantener la consistencia de los datos replicados en ambientes distribuidos, a partir de las prácticas positivas, métodos representados, y experiencias de expertos, en materia de modelos y herramientas de replicación de datos.

- La pertinencia como garantía de la adecuación del modelo en el contexto de la replicación de datos en ambientes distribuidos.

- La coherencia con las dimensiones y capacidades arquitectónicas usualmente empleadas para describir las soluciones de replicación.

- La adaptabilidad que se logra utilizando componentes con funcionalidades genéricas y sobre una estructura que facilita su aplicación a particularidades de los sistemas existentes.

-La actualización mediante la retroalimentación de la información que nutre al modelo que se logra analizando los elementos que faltan o que requieren ser adaptados cuando se identifican nuevos factores.

Las características del modelo propuesto son:

- Integralidad que comprende fundamentos teóricos, buenas prácticas, marcos de trabajo y experiencia de expertos; sobre modelos, herramientas de replicación y elementos necesarios para mantener la consistencia en ambientes distribuidos.

- Enfoque sistémico que se expresa mediante la gestión de los procesos asociados a la replicación de datos y la consistencia, de forma coherente y en conjunto.

- Mejora continua con el perfeccionamiento constante del modelo a partir de los resultados que se van obteniendo tras su aplicación en entornos de replicación.

- Amplitud que brinda la capacidad de describir múltiples procesos asociados a la replicación de datos y la posibilidad de emplearse en variados escenarios de replicación distribuidos.

**2.2 Modelo de replicación de datos propuesto**

El modelo propuesto está compuesto por cuatro componentes relacionados entre sí. Estos integran fundamentos teóricos, buenas prácticas, marcos de trabajo y experiencia de expertos, sobre modelos y herramientas de replicación.

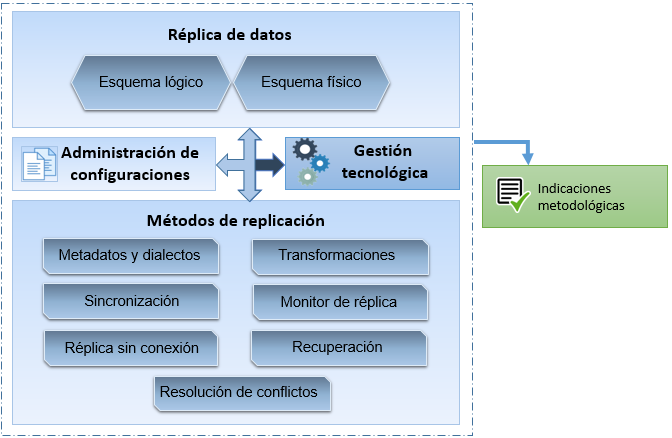


Figura 1. Concepción metodológica para elaborar el modelo. Fuente: elaboración propia.

El componente Réplica de datos describe el funcionamiento general del proceso de replicación de datos mediante la especificación de los componentes lógicos y distribución física. El componente Métodos de replicación, está diseñado para describir los procesos asociados a la réplica. Ambos componentes requieren de la Administración de configuraciones. Para la construcción de todos los componentes se emplean herramientas, estándares y servicios propuestos en el componente Gestión tecnológica. Finalmente, se plantean las indicaciones metodológicas para la implementación del modelo.

Componente Réplica de datos: el proceso de replicación propuesto está compuesto por cuatro componentes: Consola de administración responsable de realizar las configuraciones principales del software como el registro de nodos y configuración de los datos a replicar, Capturador encargado de capturar los cambios que se realizan sobre la BD, Distribuidor responsable de realizar la propagación de los cambios empleando para ello el Servidor de mensajería ActiveMQ, y el Aplicador concepto designado para aplicar los cambios en la BD.

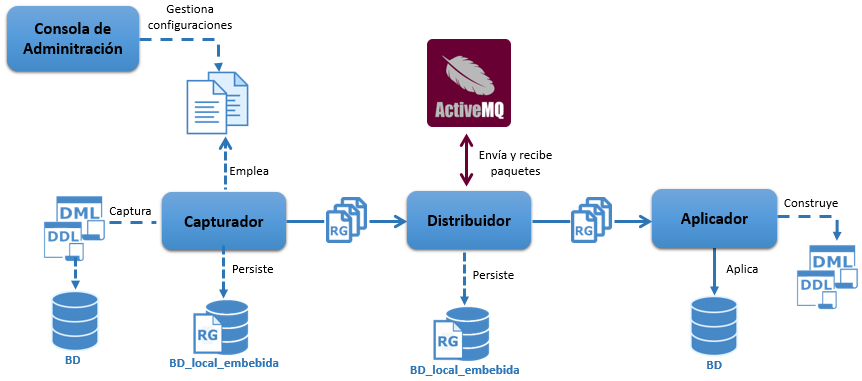
****

Figura 2. Propuesta general para la replicación de datos. Fuente: elaboración propia.

Componente Configuraciones generales: para iniciar el proceso de captura de los cambios es necesario establecer el entorno de replicación. En el entorno de replicación se deben:

- Identificar los nodos que forman parte del sistema distribuido y las conexiones a los servidores para poder establecer las comunicaciones entre las instancias, que se proponen establecer mediante configuraciones generales del sistema.

- Registrar los nodos y establecer la dirección o sentido de trasmisión de los datos, que se proponen realizar mediante la configuración de nodos.

- Especificar los cambios a replicar entre instancias que se propone condicionar mediante la configuración de réplica.

Se puede ofrecer una solución que, de acuerdo con la dirección en que fluyen los datos, brinde réplica en un solo sentido (unidireccional) o réplica en ambos sentidos (bidireccional). En ambos casos es posible determinar el sentido mediante configuraciones de réplica, donde en cada nodo se definirían las tablas a replicar y hacia qué nodo se enviarán los cambios. Además, mediante las configuraciones enunciadas anteriormente es posible realizar el proceso de réplica de forma desatendida.

Componente Gestión tecnológica: contiene una propuesta que articula el desarrollo tecnológico con el uso de herramientas libres basada en la línea de investigación JAVA. En la siguiente figura de elaboración propia si sintetizan las herramientas propuestas.



Figura 3. Componentes de una infraestructura tecnológica para réplica de datos.

**2.3 Componente Métodos de replicación**

En este componente se describen y modelan con BPMN los procesos inmersos en la replicación tales como: definición de dialectos y metadatos, sincronización, resolución de conflictos, réplica offline, transformaciones, monitor de réplica, y mecanismos de recuperación. A continuación, se sintetizan los procesos fundamentales:

Dialectos y metatados: a través de los metadatos se debe cargar la información de la BD y en los dialectos definir las sintaxis SQL correspondientes a cada gestor contemplado.

Sincronizacion: consiste en homogeneizar los datos y estructuras del nodo remoto con el nodo local según la configuración de réplica registrada. Cuando el administrador del nodo local determina que desea sincronizar con el nodo remoto, se envía una petición de sincronización hacia el nodo remoto para que autorice la modificación de su BD. La petición transita por los siguientes estados: Aceptar cuando el nodo remoto permite la sincronización, Rechazar cuando el nodo remoto se opone a la sincronización y rechaza permanentemente cualquier solicitud de ese nodo, y Cancelar cuando el nodo remoto en ese momento anula la petición.

Cuando el nodo local recibe la confirmación, se construyen los grupos replicables con los cambios y se envían a través del distribuidor de cambios, quien se responsabiliza de su propagación con el empleo del servidor de mensajería Apache ActiveMQ. Recibidos los grupos por el nodo remoto se procede a aplicar los cambios arribados, conformando acciones DML y DDL respectivamente asociadas al cambio. Si se produce un error en la BD durante la aplicación de los cambios se procede a resolver el conflicto generado según la configuración establecida, para lo cual se debe brindar la posibilidad de configurar la resolución de conflictos para cada nodo.

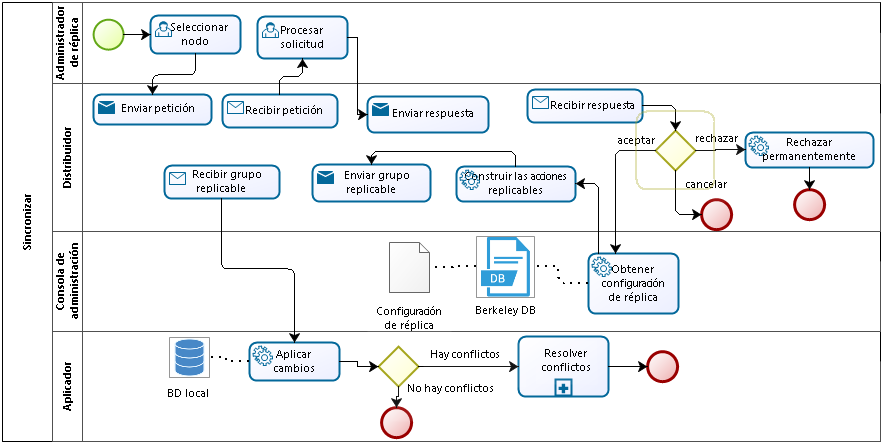


Figura 4. Proceso “Sincronizar”. Fuente: elaboración propia.

Resolución de conflictos: al aplicar los cambios en la BD, como parte de los procesos de réplica con conexión, sin conexión y sincronización, es posible que la ejecución de la sentencia DML o DDL no resulte exitosa y con ello se generen datos y/o definiciones incompletas que provocan inconsistencias entre las BDs. Por esta razón, se hace necesario diseñar un mecanismo de resolución de conflictos que lleve a la BD a un estado consistente.

La dinámica del proceso consiste en capturar el conflicto cuando se produce el error al aplicar la sentencia DML o DDL en la BD, luego se clasifica y finalmente se resuelve según el tipo de conflicto y variante configurada para la resolución en las configuraciones generales. El conflicto se clasifica mediante el SQL State que genera la BD cuando se produce el error al aplicar el SQL y las variantes determinan la manera de resolver los conflictos. Para definir las variantes de solución al conflicto capturado, se consideraron las siguientes condiciones:

* Si se define el nodo remoto como Ganador entonces la información recibida que entra en conflicto tiene prioridad y prevalece sobre la existente en el nodo local.
* Si se define el nodo remoto como No Ganador entonces la información recibida que entra en conflicto no tiene prioridad sobre la existente en el nodo local.
* En cualquier variante de solución que sea aplicada, debe mantenerse la consistencia entre las BDs.

Cada vez que sea capturado un conflicto debe ser resuelto. Si el nodo local no está en conflicto y no tiene grupos pendientes por ser aplicados se procederá a resolverlo, de lo contrario será persistido en una BD embebida para esperar a que pueda ser aplicado. Si la resolución del conflicto es automática se resolverá automáticamente según la variante que se haya definido en las configuraciones generales. Si la configuración definida es manual, en el momento que se genere el conflicto el usuario podrá decidir qué variante aplicar para dar solución al mismo, teniendo en cuenta las variantes definidas para el tipo de conflicto generado. Cuando el nodo local resuelva el conflicto serán monitorizadas las acciones de las instancias involucradas con el conflicto.

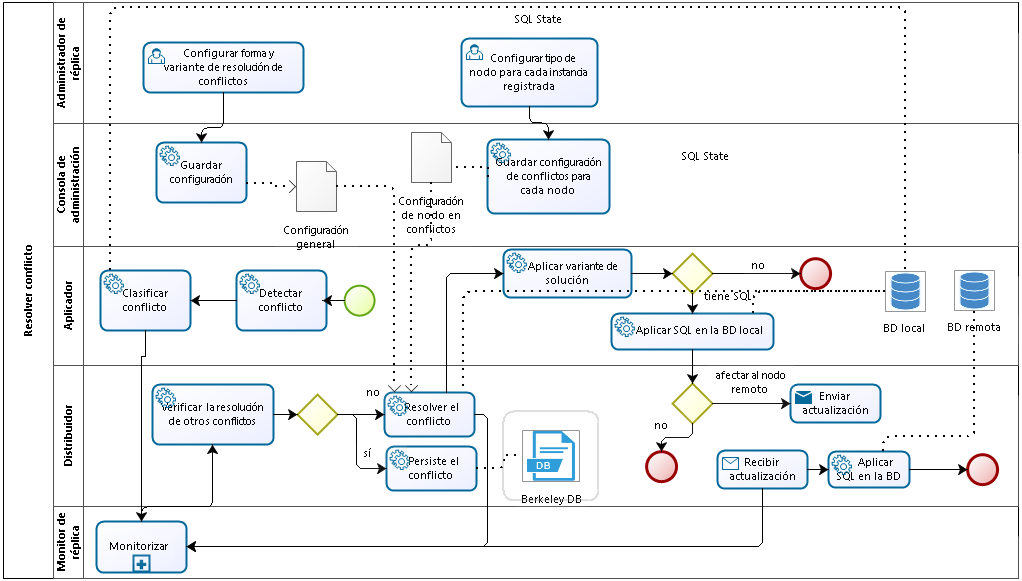


Figura 5. Proceso “Resolver conflictos”. Fuente: elaboración propia.

**3. Resultados y discusión**

Para validar la propuesta se aplicaron métodos cuantitativos y cualitativos que posteriormente fueron triangulados para lograr una mayor precisión y objetividad de las comprobaciones. El escalamiento de Likert para validar los componentes del modelo. La técnica Iadov se aplica para medir el índice de satisfacción de los usuarios con respecto al modelo propuesto. Grupo focal para definir las medidas para los indicadores definidos en las dimensiones de la variable dependiente, con el propósito de evaluar la consistencia de los datos; y conocer además la opinión de los expertos sobre los beneficios y facilidades de uso del modelo propuesto.

Inicialmente se realizó una valoración inicial de los expertos para validar el modelo. En el proceso de selección se consultaron especialistas con experiencia en tecnologías de gestión de datos, desarrollo y/o despliegue de soluciones de replicación de datos, pertenecientes a los centros CDAE, DATEC, DIN y FORTES. Se evaluaron sus conocimientos sobre su experiencia y conocimiento en el tema de replicación de datos, investigaciones y/o publicaciones, así como actualización en cursos de postgrado, diplomados, maestrías, entre otros como la vinculación como tutores en trabajos de diplomas. A partir de esta valoración se seleccionaron los expertos para aplicar cada técnica.

**3.1 Validación de la consistencia en un entorno de réplica**

Se realizó el proceso de instrumentación aplicando la propuesta en un escenario de replicación, mediante pruebas funcionales y las métricas definidas por expertos del grupo focal. Como estrategia de pruebas se efectuaron pruebas al sistema. Específicamente se realizaron pruebas funcionales aplicando la técnica de Caja Negra con el método de Partición equivalente. Para lo cual se emplearon las BDs embebidas, los datos mostrados por el Monitor de réplica, entre otras opciones del Replicador de datos REKO que constituye la solución informática resultante del aporte práctico de la presente investigación. Las métricas definidas para cada escenario son las siguientes:

Escenario 1. Evaluar la integridad referencial, completitud y coherencia de los datos: , donde A representa la cantidad de acciones aplicadas y R la cantidad de acciones replicadas.

Escenario 2. Evaluar la homogeneidad de las bases de datos en la sincronización: , donde C representa la cantidad de acciones construidas para homogeneizar las BDs según la configuración de réplica registrada.

Escenario 3. Evaluar la capacidad de resolución de conflictos: , donde R representa la cantidad de conflictos resueltos y G la cantidad de conflictos generados.

Escenario 4. Evaluar la persistencia ante eventualidades: , donde E representa la cantidad de grupos enviados que se encontraban pendientes y P la cantidad de grupos persistidos en el origen pendientes por ser enviados.

Escenario 5. Evaluar la capacidad de operar sin conexión: , donde M representa la cantidad de acciones importadas que se aplicaron en el destino y X la cantidad de acciones exportadas en los archivos en el origen.

Para analizar los resultados, durante el desarrollo de la guía del grupo focal, expertos en el tema definieron D como el conjunto formado por los indicadores correspondientes a las dimensiones que pueden establecerse para evaluar la consistencia de los datos (). Entonces, para que la consistencia de los datos sea correcta según los indicadores debe cumplirse que:

(1)

La propiedad anterior expresa que los indicadores en todas las dimensiones establecidas deben arrojar un valor igual a 1. De lo contrario, si existe al menos diferente de 1 el resultado es incorrecto, y por ende se dice que el sistema no mantiene la consistencia.

Por tanto, los resultados mostrados en los cinco escenarios fueron satisfactorios, debido a que el indicador para mediar la consistencia (I) fue igual a 1 en todos los casos, con lo cual se corrobora que el replicador de datos REKO, mantiene la consistencia de los datos replicados, y por ende se ratifica la contribución del modelo para mantener la consistencia de los datos en ambientes distribuidos. A continuación, se visualizan los resultados de dos de los escenarios validados. Aunque, cabe señalar que todos los escenarios fueron validados y en todos se obtuvieron resultados satisfactorios.

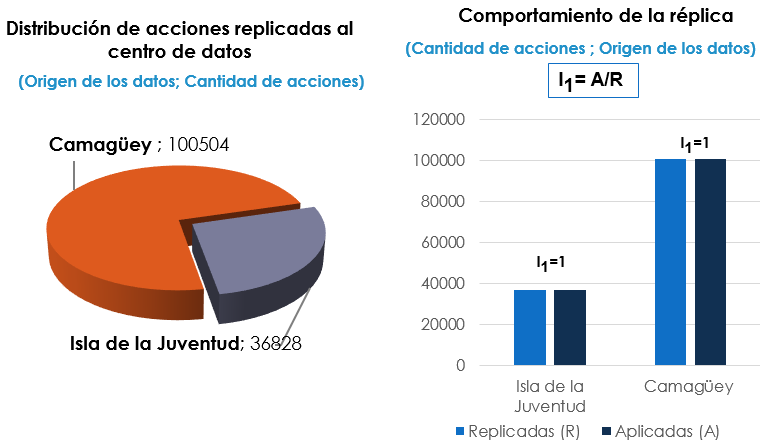
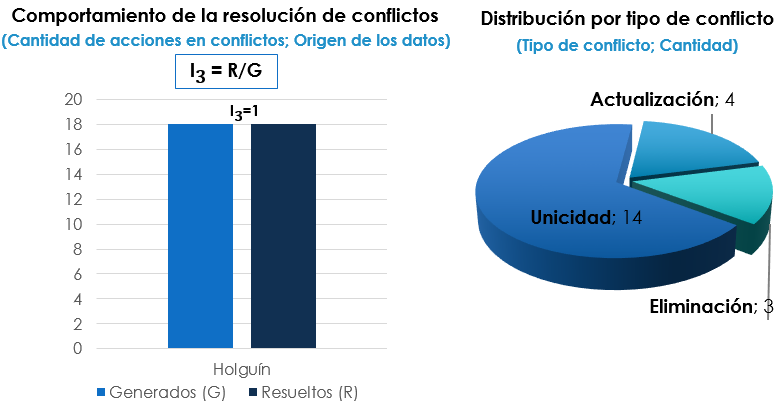
 

Figura 6. Resultados del Escenario 1. Figura 7. Resultados del Escenario 3.

**3.2 Validación de la satisfacción de los usuarios al aplicar la técnica Iadov**

El 91% de los encuestados tiene una clara satisfacción, el 8% está más satisfecho que insatisfecho. Estos resultados representan un Índice de Satisfacción Grupal de 0.95. Este valor confirma resultados favorables para la aceptación de la propuesta desarrollada, pues se encuentra dentro de la categoría grupal, de satisfacción de 0.5 a 1. Lo que significa una clara satisfacción con la propuesta y reconocimiento de su utilidad para contribuir a mantener la consistencia de los datos en ambientes distribuidos. Para demostrar la confiabilidad de las respuestas dadas por los expertos en los cuestionarios se realizó el cálculo del coeficiente de confiabilidad Alfa-Cronbach sobre la base de la varianza de los ítems. Se evidencia un buen grado de confiabilidad, al obtener un coeficiente de 0,84.

Figura 8. Resultados de la aplicación de la técnica Iadov y valor del índice de satisfacción grupal. Fuente: elaboración propia.

**3.3 Valoración de los expertos al aplicar la escala psicométrica Likert**

Al aplicar la escala psicométrica Likert, se obtuvo un índice porcentual superior a 80% en todos los casos, lo cual evidenció que tanto los elementos teóricos como los principios, características y componentes del modelo, tienen una alta valoración por parte de los expertos. Así como la comprensión, pertinencia, coherencia y adaptabilidad del modelo propuesto.

Figura 8. Resultados de la aplicación de la escala Likert. Fuente: elaboración propia

**4. Conclusiones**

En el modelo propuesto se describen y modelan con BPMN todos los procesos inmersos en la replicación identificados en la literatura, abordando en cada uno la consistencia de los datos, siendo esta la diferencia, aporte y novedad con respecto a otros modelos. Además de describirse "qué" hacer, se fundamenta el "cómo" hacerlo y se proponen tecnologías adecuadas de software libre, con lo cual se superan estas dificultades identificadas en los demás modelos y soluciones examinadas en la literatura.

El análisis de los referentes teóricos relacionados con los modelos, procesos y soluciones de replicación de datos, permitió analizar las tendencias actuales en el desarrollo de soluciones de réplica e identificar un grupo de insuficiencias, que atentan contra la capacidad de los modelos para contribuir a mantener la consistencia en la replicación de datos.

El desarrollo de la propuesta conformada por cuatro componentes y la descripción de los procesos asociados acordes a buenas prácticas, métodos representados y experiencias de expertos, en materia de modelos y herramientas de replicación de datos, posibilitó organizar y estructurar un modelo para la réplica de datos que contribuye a mantener la consistencia de los datos replicados.

La validación del modelo mediante la aplicación de métodos científicos y pruebas funcionales en un entorno real, corroboró que los componentes del mismo están alineados a las tecnologías más actuales; y existen opiniones favorables y una alta satisfacción de usuarios potenciales con respecto a la aplicabilidad, actualidad y factibilidad de uso, del modelo. La instrumentación del modelo mediante la herramienta desarrollada y las pruebas realizadas demostraron su contribución para mantener la consistencia de los datos replicados en ambientes distribuidos.

**5. Referencias bibliográficas**

[1]. PowerData. ¿Qué se entiende por integridad de los datos? Especialistas en Gestión de datos, 2013. https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/348870.

[2]. Silberschatz, A.; Korth, H. F.; y Sudarshan, S. Fundamentos de Bases de Datos. 4ta ed. España, McGraw-Hill, 2002.

[3]. Vila, Joan. Replicación en sistemas distribuidos. DISCA/UPV Valencia, 2017.

[4]. Escobar, A. Replicación multimaestro en BD PostgreSQL. España 2017, p. 9-10.

[5]. Len, S. J. Base de Datos Distribuidas Estudio de Actualización de Réplicas. Facultad de Informática, Argentina, 2001.

[6]. Pérez, J. L. Modelo de replicación de datos en sistemas distribuidos de bases de datos relacionales. Universitas Major Pacensis Divi Andreae, 40-89p, Bolivia. 2006.

[7]. Erazo, Roberto; Cobos, Carlos y Mendoza, Martha. Modelo de servicios web para replicación de instantáneas. Revista Facultad de Ingeniería 2014, p. 144–157.

[8]. Microsoft Corporation. Sincronizar y resolver conflictos en un conjunto de réplicas (MDB). 2014. http://office.microsoft.com/.

[9]. Oracle Help Center. Streams Concepts and Administration. 2018. https://docs.oracle.com/cd/E11882\_01/server.112/e17069.pdf.

[10]. PostgreSQL. Announcing the release of pglogical 2.0. Oxford, United Kingdom 2017. https://www.postgresql.org/about/news/1744/.

[11]. Community SymmetricDS. 2018. http://www.symmetricds.org/about/overview

[12]. Continuent. Products and Solutions. 2018. https://www.continuent.com/solutions/