XX JORNADAS DE INGENIERÍA DEL SOFTWARE Y BASES DE DATOS

Santander, del 15 al 17 de Septiembr
de 2015

PRESENTACIÓN

Las JISBD son un foro de referencia en la investigación de la Ingeniería del Software y las Bases de Datos e ámbito iberoamericano. A lo largo de los años, el evento ha servido para que los investigadores de Españ Portugal y Latinoamérica presentasen sus trabajos y establecieran una comunidad muy sólida alrededor de a disciplinas. En 2015, las Jornadas celebran su XX edición. Es, por ello, una ocasión para hacer balance del ca recorrido, por un lado, y de consolidar el papel dinamizador de la comunidad a la que aloja, por otro.

Es un hecho incontestable que la comunidad de JISBD ha crecido considerablemente desde el inicio de la Jornadas. Con el paso del tiempo, los grupos han abierto nuevas líneas de investigación, por lo que, al ampa JISBD han surgido un conjunto de comunidades cuyo interés abarca partes específicas de ambas disciplin Atendiendo a esa diversidad, y con el fin de que las Jornadas sean un punto de encuentro entre los miembro cada comunidad, y de ellas entre sí, el programa de las JISBD 2015 va a estructurarse en Sesiones Temátic "tracks". Cada una de ellas se ha organizado alrededor de una comunidad científica que comparte interés p ciertos temas de investigación en Ingeniería del Software. Con el fin de acoger a trabajos que no tengan acor temático en ninguno de los tracks seleccionados, se ha incluido un track abierto con una lista de temas amp

JISBD 2015 es una conferencia organizada bajo los auspicios de Sistedes (http://www.sistedes.es (Sociedad de Ingeniería del Software y Tecnologías de Desarrollo de Software).

XX Jornadas de Ingeniería del Software y Bases de Datos

PROGRAMA

Programas de las Jornadas

Descargar el programa completo (Programa_JISBD.pdf)

ACTAS

KEYNOTE

Evaluating Research (And Researchers): Should We Care? -A Software Engineering Perspective
(Actas\JISBD\abstractCarloGhezzi.pdf)
Carlo Ghezzi.

SESIÓN DSDM1: DESARROLLO DE SOFTWARE DIRIGIDO POR MODELOS.

CEViNEdit: mejorando el proceso de creación y personalización de editores gráficos cognitivamen eficaces con GMF (Actas\JISBD\5_DSDM\01_paper_61.pdf)
David Granada, Ángel Moreno, Juan Manuel Vara, Veronica Andrea Bollati and Esperanza Marcos.

Automatización para la edición de modelos basada en vistas de dominio (Actas\JISBD\5_DSDM \02_paper_7.pdf)
César Cuevas, Patricia López Martínez and Jose M. Drake.

PyEmofUC: Un entorno MDE/EMOF minimalista (Actas\JISBD\5_DSDM\03_paper_10.pdf)
José M. Drake, César Cuevas, Juan Ramón Fernández Castañera and Patricia López Martínez.

SESIÓN DSDM2: DESARROLLO DE SOFTWARE DIRIGIDO POR MODELOS.

Model Driven NoSQL Data Engineering (Actas\JISBD\5_DSDM\04_paper_67.pdf)
Diego Sevilla, Severino Feliciano Morales and Jesus Garcia-Molina.

Achieving software-assisted knowledge generation through model-driven interoperability (Actas\JIS \5_DSDM\05_paper_11.pdf)
Patricia Martin-Rodilla, Giovanni Giachetti and Cesar Gonzalez-Perez.

Mediación semántica A* basada enMDE para la generación de arquitecturas en tiempo de ejecució
Javier Criado, Luis Iribarne and Nicolás Padilla.

Lenguaje de Modelado para Escenarios de Inteligencia Ambiental.
ablo Campillo-Sánhcez, Juan Pavón and Jorge J. Gómez-Sanz.

SESIÓN DSDM3: DESARROLLO DE SOFTWARE DIRIGIDO

POR MODELOS.

Aplicando DSDM al Diseño, Implementación y Verificación de Software para Drones: Una Primera Aproximación. (Actas\JISBD\5_DSDM\08_paper_63.pdf)

Enrique Moguel, Cristina Vicente-Chicote and Juan Hernández.

Lenguaje específico del dominio para generación de aplicaciones de procesos administrativos. (Actas\JISBD\5_DSDM\09_paper_3.pdf)

Antonio García Domínguez, Ismael Jerez Ibáñez and Inmaculada Medina Bulo.

Arquitectura basada en modelos para la generación de especificaciones textuales de requisitos a part procesos de negocio definidos mediante BPMN. (Actas\JISBD\5_DSDM\10_paper_22.pdf)

José Manuel Cruz Zapata, Begoña Moros Valle and José Ambrosio Toval Álvarez.

Analysis of the Scientific Production of the Spanish Software Engineering Community. (Actas\JISB \5_DSDM\11_paper_33.pdf)

Loli Burgueño, Antonio Moreno-Delgado and Antonio Vallecillo.

SESIÓN GD1: GESTIÓN DE DATOS

Un índice espacio-temporal compacto para consultas time-slice y time-interval. (Actas\JISBD \1_GD\1_paper_43.pdf)

Nieves R. Brisaboa, Ramón Casares, Andrea Rodríguez, Miguel Romero and Diego Seco.

Query approximation in the case of incompletely aligned datasets. (Actas\JISBD\1_GD\2_paper_54.p

Ana Isabel Torre Bastida, Jesús Bermúdez and Arantza Illarramendi.

Modernizing secure OLAP applications with a model driven approach.

Carlos Blanco, Eduardo Fernandez-Medina and Juan Trujillo.

A First Step Towards Keyword-Based Searching for Recommendation Systems. (Actas\JISBD \1_GD\4_paper_28.pdf)

María Del Carmen Rodríguez-Hernández, Francesco Guerra, Sergio Ilarri and Raquel Trillo Lado.

SESIÓN GD2: GESTIÓN DE DATOS

Integración semántica de datos de observación mediante servicios SOS. (Actas\JISBD\1_GD\5_paper_3

Manuel A. Regueiro, José R.R. Viqueira, Christoph Stasch and José Cotos.

Optimización del Almacenamiento de Datos en la Gestión Energética de Edificios Inteligentes. (Actas\J \1_GD\6_paper_21.pdf)

Samuel Otero Paz, Jose Angel Taboada, José R.R. Viqueira and Juan Enrique Arias Rodriguez.

Un marco para democratizar la minería de datos: propuesta inicial y retos. (Actas\JISBD \1_GD\7_paper_25.pdf)

Diego García-Saiz, Roberto Espinosa, José Jacobo Zubcoff, José-Norberto Mazón and Marta Zorrilla.

SESIÓN GD3: GESTIÓN DE DATOS

SODA: A framework for spatial observation data analysis.

Sebastián Villarroya, José R.R. Viqueira, Manuel A. Regueiro, Jose Angel Taboada and Jose Cotos

Tracing Conceptual Models' Evolution in Data Warehouses by using the Model Driven Architecture

Alejandro Maté and Juan Trujillo.

Efficient XPath Evaluation on Compressed XML Documents.

Nieves R. Brisaboa, Ana Cerdeira-Pena and Gonzalo Navarro.

Compressed vertical partitioning for efficient RDF management.

Sandra Álvarez-García, Nieves R. Brisaboa, Javier D. Fernández, Miguel A. Martínez-Prieto and Gonzalo Nav

SESIÓN IWSC: INGENIERÍA WEB Y SISTEMAS COLABORATIVOS

JET: A Proof of Concept Enabling Mobile Devices as Personal Profile Providers. (Actas\JISBD\2_IWS\1_paper_27.pdf)

Javier Berrocal, Carlos Canal, Jose García-Alonso, Niko Mäkitalo, Tommi Mikkonen, Javier Miranda and Ju Manuel Murillo Rodríguez.

GeoNews: Generación automática de contextos geográficos para programas de noticias a través de H (Actas\JISBD\2_IWSC\2_paper_30.pdf)

Moisés Vilar, Sebastián Villarroya, José R.R. Viqueira and Jose Cotos.

Modelos de Contexto en el Desarrollo de Interfaces Post-WIMP: una Revisión Crítica. (Actas\JISBD\2_I\3_paper_60.pdf)

Arturo C. Rodríguez, Cristina Roda, Elena Navarro and Pascual González.

Empirical study on the maintainability of Web applications: Model-driven Engineering vs Code-centr

Yulkeidi Martínez, Cristina Cachero and Santiago Meliá.

SESIÓN PSM: PROCESOS SOFTWARE Y METODOLOGÍA

A View of Process Improvement from an Academic Perspective: How Does Software Engineering Educ Contribute to CMMI Practices?

Ana M Moreno, Maribel Sánchez-Segura, Fuensanta Medina-Domínguez and Gonzalo Cuevas.

On the Impact of UML Analysis Models on Source-Code Comprehensibility and Modifiability

Giuseppe Scanniello, Carmine Gravino, Marcela Genero, José A. Cruz-Lemus and Genny Tortor.

SESIÓN CP: CALIDAD Y PRUEBAS

Pruebas basadas en flujo de datos para programas MapReduce. (Actas\JISBD\4_CP\1_paper_36.pdf

Jesús Morán, Claudio De La Riva and Javier Tuya.

I8K|DQ-BigData: Extensión Arquitectura I8K para Calidad de Datos en Big Data. (Actas\JISBD\4_CP\2_paper_73.pdf)

Bibiano Rivas, Jorge Merino, Manuel Serrano, Ismael Caballero and Mario Piattini.

Automated metamorphic testing of variability analysis tools.

Sergio Segura, Amador Duran, Ana B. Sánchez, Daniel Le Berre, Emmanuel Lonca and Antonio Ruiz-Cort

Coverage-based testing for Service Level Agreements.

Marcos Palacios, José García-Fanjul, Javier Tuya and George Spanoudakis.

Herramienta para la Prueba de Mutaciones en el Lenguaje C++. (Actas\JISBD\4_CP\5_paper_34.pdf)

Pedro Delgado-Pérez, Inmaculada Medina-Bulo and Juan José Domínguez-Jiménez.

SESIÓN ASV1: ARQUITECTURA SOFTWARE Y VARIABILID

Desarrollo de una Línea de Productos Software utilizando las clases parciales C#: Slicer Pattern.

(Actas\JISBD\6_ASV\1_paper_41.pdf)

Alejandro Pérez Ruiz and Pablo Sánchez Barreiro.

Defining and Validating a Feature-Driven Requirements Engineering Approach.

Raphael Pereira de Oliveira, David Blanes, Javier Gonzalez-Huerta, Emilio Insfran, Silvia Abrahao, Sholom C and Eduardo Santana de Almeida

A model for tracing variability from features to product-line architectures: a case study in smart gri

Jessica Díaz, Jennifer Perez and Juan Garbajosa.

SESIÓN ASV2: ARQUITECTURA SOFTWARE Y VARIABILID

Propuesta para un acceso homogéneo a servicios PaaS en la Nube (Actas\JISBD\6_ASV\4_paper_41.

Miguel Barrientos, Jose Carrasco Mora, Javier Cubo and Ernesto Pimentel.

Exploring the Synergies between Joing Point Interfaces and Feature-Oriented Programming. (Actas\JI \6_ASV\5_paper_39.pdf)

Cristian Vidal Silva, David Benavides, José Galindo and Paul Leger.

SESIÓN SBSE1: INGENIERÍA DEL SOFTWARE GUIADA PO BÚSQUEDAS

Exact Scalable Sensitivity Analysis for the Next Release Problem

Mark Harman, Jens Krinke, Inmaculada Medina Bulo, Francisco Palomo Lozano, Jian Ren and Shin Yoo

Análisis de las soluciones guiadas por búsqueda para el problema de selección de requisitos. (Actas\J \7_SBSE\2_paper_23.pdf)

Isabel Del Águila and José Del Sagrado.

Resolviendo un problema multi-objetivo de selección de requisitos mediante resolutores del problema

(Actas\JISBD\7_SBSE\3_paper_42.pdf)

Isabel Del Águila, José Del Sagrado, Francisco Chicano and Enrique Alba.

TESTAR - Automated User Interface Testing Tool for Industry Adoption (Actas\JISBD\7_SBSE \4_paper_71.pdf)

Urko Rueda, Tanja E.J. Vos, Francisco Almenar, Mirella Oreto and Anna Esparcia

SESIÓN SBSE2: INGENIERÍA DEL SOFTWARE GUIADA PO BÚSQUEDAS

QoS-aware web services composition using GRASP with Path Relinking

José Antonio Parejo Maestre, Sergio Segura, Pablo Fernandez and Antonio Ruiz Cortés

People as a Service y la Ingeniería del Software Guiada por Búsqueda (Actas\JISBD\7_SBSE\6_paper_2

Jose García-Alonso, Jose Javier Berrocal Olmeda and Juan Manuel Murillo Rodríguez

Automated generation of computationally hard feature models using evolutionary algorithms

Sergio Segura, José Antonio Parejo Maestre, Rob Hierons, David Benavides and Antonio Ruiz-Cortés

SESIÓN SBSE3: INGENIERÍA DEL SOFTWARE GUIADA PO BÚSQUEDAS

Interactividad en el descubrimiento evolutivo de arquitecturas software (Actas\JISBD\7_SBSE \8_paper_44.pdf)

Aurora Ramírez, José Raúl Romero and Sebastián Ventura

Análisis y determinación del impacto del operador de mutación en la generación genética de casos prueba para WS-BPEL. (Actas\JISBD\7_SBSE\9_paper_40.pdf)

Antonia Estero-Botaro, Álvaro Cortijo-García, Antonio García-Domínguez, Francisco Palomo-Lozano, Juan J Domínguez-Jiménez and Inmaculada Medina-Bulo.

SESIÓN OPEN

SHERLOCK: Semantic management of Location-Based Services in wireless environments

Roberto Yus, Eduardo Mena, Sergio Ilarri and Arantza Illarramendi

Understanding replication of experiments in software engineering: A classification

Omar S. Gómez, Natalia Juristo and Sira Vegas

Un entorno de gestión de casos para la resolución flexible de emergencias (Actas\JISBD\8_OPEN \3_paper_58.pdf)

Juan Sánchez, José Carsí and Carmen Penadés

KVLEAP: Interacción sin contacto (touchless) con ordenadores (Actas\JISBD\8_OPEN\4_paper_56.p

Kevin Villalobos, David Anton, Alfredo Goñi and Arantza Illarramendi

(https://www.linkedin.com

/groups

(https://sistedes.

/Sistedes-

/sistedes)

4264422)

I8K|DQ-BigData: Extensión Arquitectura I8K para Calidad de Datos en Big Data

Bibiano Rivas, Jorge Merino, Manuel Serrano, Ismael Caballero, Mario Piattini

Instituto de Tecnologías y Sistemas de Información

Universidad de Castilla–La Mancha.

Camino de Moledores s/n, 13071, Ciudad Real

{Bibiano.Rivas, Jorge.Merino, Manuel.Serrano, Ismael.Caballero, Mario.Piattini}@uclm.es

abstract.
Durante la ejecución de procesos de negocios que implican a varias organizaciones, normalmente se intercambian Datos Maestros. Es necesario que dichos datos tengan niveles adecuados de calidad, ya que de otro modo, puede ocurrir que los procesos de negocio fallen. Si los datos intercambiados llevasen información sobre su nivel de calidad, entonces sería posible decidir si pueden usarse o no en dichos procesos. Las partes 100 a 140 de ISO/TS 8000 pueden ayudar a proporcionar esta información de forma usable. En concreto I8K, una implementación de referencia con fines académicos del citado estándar, puede ser usado para tal fin. Lamentablemente, la eficiencia de I8K cae cuando se trata de evaluar la calidad de grandes volúmenes de Datos Maestros. Este artículo describe la extensión realizada sobre la arquitectura I8K para solventar los problemas de rendimiento al evaluar grandes volúmenes de datos utilizando tecnologías Big Data.

Keywords: Big Data - Calidad de Datos - I8K – Intercambio de datos Maestros.

1 Introducción

Cada día se informatiza, automatiza y se procesa un número cada vez mayor de datos llegando a volúmenes hace décadas inimaginables. Estos datos son habitualmente almacenados e intercambiados entre organizaciones [1]. Los datos son uno de los activos más importantes de una organización. Para poder tener el mayor beneficio de dichos datos, es necesario que tengan un nivel de calidad adecuado para las tareas para la que

se quieren utilizar. Si esto no ocurre, puede que las operaciones y procesos de toma de decisiones lleguen a fracasar o no consigan cumplir sus objetivos [2]. En este sentido las organizaciones podrían beneficiarse del hecho de poder adherir información, junto a los datos intercambiados, sobre el nivel de calidad que tienen los datos que están siendo intercambiados. Si esta información estuviera presente, se podría tener en cuenta a la hora de incluir ciertos datos en ciertas operaciones de los procesos de negocio, o incluso, se puede exigir que se intercambien datos que tengan un cierto nivel de calidad mínimo. Los datos que normalmente se intercambian son los llamados Datos Maestros, que son aquellos que representan los conceptos básicos de las organizaciones que se incluyen en los procesos de negocio. Así, Loshin define Datos Maestros como "*aquellos objetos esenciales para el negocio usados en las diferentes aplicaciones de una organización, junto con sus metadatos asociados, atributos, definiciones, roles, conexiones y taxonomías*."[3].

Las partes 100 a 140 de ISO 8000 [4-8] describen los requisitos que se tienen que satisfacer para poder asegurar el nivel de calidad de datos en el intercambio de Datos Maestros [9]. Estos requisitos han sido implementados en una arquitectura de servicio denominada I8K, descrita por Caballero et al en [10]. Junto a la arquitectura de servicio, se incluye ICS-API, una interfaz de programación de aplicaciones que permite a los desarrolladores de aplicaciones explotar los servicios de I8K cuando tienen que recurrir a la evaluación de la calidad de los datos intercambiados.

No obstante, las pruebas realizadas usando I8K para grandes volúmenes de datos revelan problemas de rendimiento debidos a la tecnología utilizada, que en esas circunstancias, raya el límite de sus prestaciones cuando se trata de procesar conjunto de datos en torno a varias decenas de megabytes. En este sentido, podría entenderse que el problema se puede afrontar desde una perspectiva Big Data.

Gartner define Big Data como datos que tienen "*gran volumen, gran velocidad y/o gran variedad de conjuntos de datos que requieren nuevas formas de procesamiento para permitir una toma de decisiones mejorada, percepción de descubrimiento y optimización de procesos*"[11]. Por tanto, se tiene un problema de Big Data a la hora de evaluar la calidad de grandes volúmenes de datos, lo que requiere la utilización de tecnologías específicas de Big Data dentro del núcleo de I8K.

En concreto, y como principal aportación dentro de este artículo, se explicará cómo se ha extendido I8K con la capacidad de evaluación de calidad de grandes volúmenes de datos utilizando un entorno Big Data, dando lugar a un nuevo producto que hemos llamado I8K-BiDa. Adicionalmente, en este artículo se explica cómo se han reescrito los algoritmos, utilizando Hadoop [12] y el modelo de programación Map-Reduce [13], para evaluar las dimensiones de calidad de datos de Precisión y Completitud propuestas en las partes 130 y 140 de ISO/TS 8000, respectivamente.

El resto del artículo está estructurado como sigue: la sección 2 resume brevísimamente el estado del arte, específicamente, la implementación de referencia I8K. En la sección

3, se propone una arquitectura Big Data con el objetivo de incorporar los conceptos de las soluciones Big Data en I8K. En la sección 4, se lleva a cabo un ejemplo de aplicación. En la sección 5 se exponen las conclusiones y trabajos futuros.

2 Trabajos relacionados

2.1. Gestión de calidad en el intercambio de Datos Maestros: ISO 8000-1x0 e I8K e ICS-API

En [14] se definen Datos Maestros como aquellos conceptos que suponen el conocimiento básico del dominio en el que una organización desarrolla su actividad. Así, las organizaciones que necesitan intercambiar Datos Maestros para la ejecución de sus procesos de negocio, deberían hacer referencia a conceptos análogos representados por versiones coherentes de Datos Maestros. Además, es importante también gestionar adecuadamente la calidad de los valores de los datos maestros que son intercambiados.

ISO/TS 8000, en concreto las partes 100 a 140, describen una serie de requisitos que permite gestionar la calidad de los datos en el intercambio de Datos Maestros entre organizaciones.

- **ISO 8000:100**: describe los aspectos específicos de Datos Maestros para gestionarlos en sistemas de Gestión de Calidad de Datos.
- **ISO 8000:102**: describe el vocabulario relacionado con la Calidad de Datos Maestros referente en las distintas partes del estándar.
- **ISO 8000:110**: establece los términos en los que se codifican los Mensajes de Datos Maestros (aspectos sintácticos, codificación semántica y requisitos).
- **ISO 8000:120**: establece describe requisitos para la representación y el intercambio de información sobre la procedencia de los Datos Maestros (Data Provenance).
- **ISO 8000:130**: establece cómo añadir información sobre el grado de *precisión* que tienen los datos.
- **ISO 8000:140**: establece cómo añadir información sobre el grado de *compleción* que tienen los datos.
- **ISO 8000:150**: establece los principios fundamentales de Gestión de Datos Maestros, así como los requisitos de implementación, intercambio y procedencia de datos.

En [15] se propone I8K una arquitectura orientada a servicio como una implementación de ISO/TS 8000, partes 100 a 140, para el intercambio de Datos Maestros entre organizaciones. Como valor extra, I8K fue concebido pensando en que podría ser un agente homologado para certificar los niveles de calidad de datos. La arquitectura de I8K para datos regulares se muestra en la **Fig. 1**.

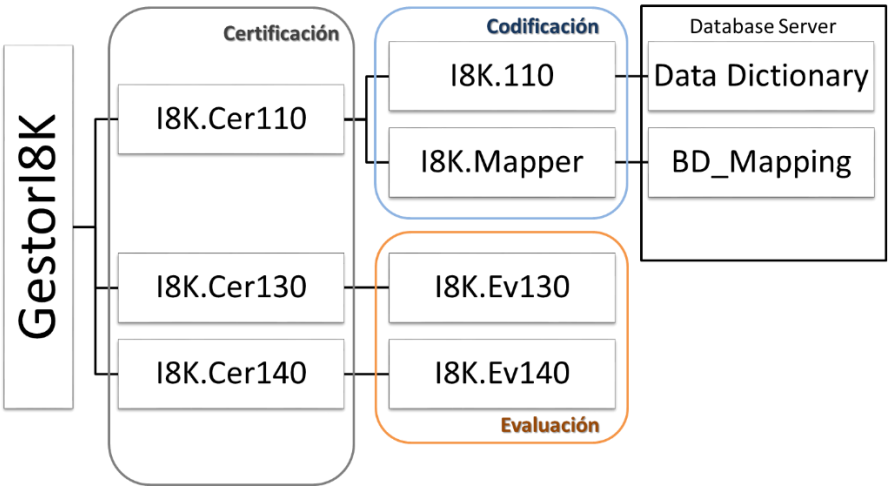

Fig. 1. Arquitectura I8K

La arquitectura de I8K está formada por los siguientes módulos:

- **Agente GestorI8K**: gestiona las peticiones que se realizan a la Arquitectura I8K, delegando en los correspondientes Agentes.
- **Agente I8K.Ev130**: evalúa la Precisión de los Datos Maestros del Mensaje de Datos Maestros.
- **Agente I8K.Cer130**: añade información de certificación de la Precisión de los Datos Maestros del Mensaje de Datos Maestros.
- **Agente I8K.Ev140**: evalúa la Completitud de los Datos Maestros del Mensaje de Datos Maestros.
- **Agente I8K.Cer140**: añade información de certificación de la Completitud de los Datos Maestros del Mensaje de Datos Maestros.

Para facilitar la comunicación de las aplicaciones que intercambian datos con I8K por parte de las aplicaciones, los autores desarrollaron una interfaz de programación de aplicaciones (API) denominada (ICS-API). La forma en la que I8K e ICS-API se utilizan se muestra en **Fig. 2**.

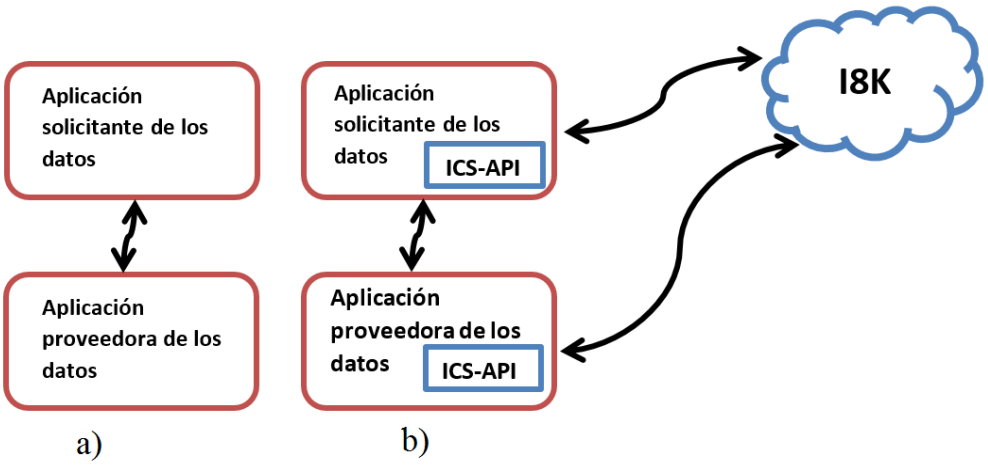

Fig. 2. a) Forma en la que dos aplicaciones interactúan sin I8K; b) Forma en la que dos aplicaciones interactúan utilizando los servicios de I8K

I8K implementa, en Java 7, los requisitos de ISO 8000-1x0 mediante servicios Web SOA. Para dar soporte a la operativa de I8K se han identificado una serie de Mensajes de Datos Maestros que son intercambiados. La secuencia de mensajes regulados por el protocolo de comunicación se muestra en la **Fig. 3.** En la **Tabla 11** se listan los tipos de mensajes que se intercambian las aplicaciones con I8K y en la **Tabla 2** se listan los mensajes entre aplicaciones que intercambian los Datos Maestros.

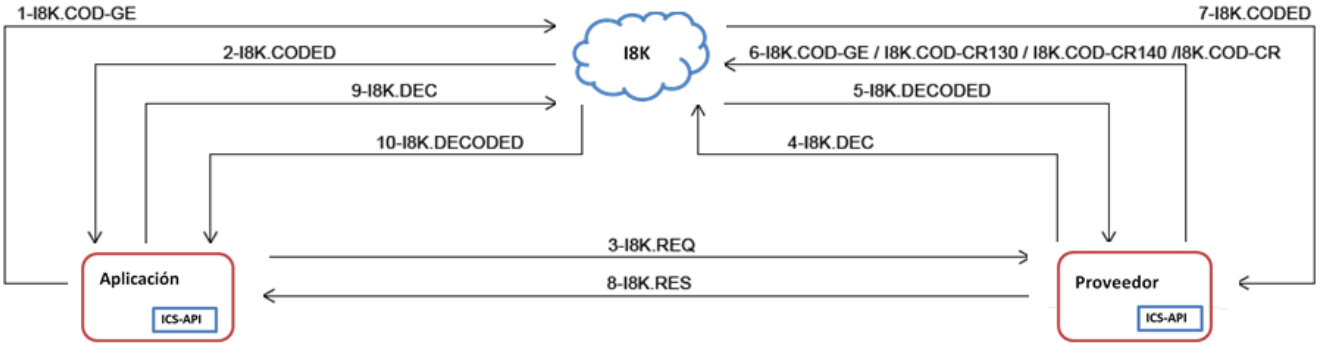

Fig. 3. Protocolo de comunicación.

Tipo	Descripción
I8K.COD-GE	Una aplicación necesita codificar datos maestros para realizar una petición a un servicio.
I8K.CODED	I8K ha codificado un mensaje de datos maestros y el contenido es devuelto a la aplicación solicitante.
I8K.DEC	Una aplicación necesita decodificar un mensaje de datos maestros recibido para entender el contenido.
I8K.DECODED	I8K ha decodificado un mensaje de datos maestros y el contenido es devuelto a la aplicación solicitante.
I8K.COD-CR130	Una aplicación necesita codificar el mensaje, evaluar y certificar la *Precisión* de los datos.
I8K.COD-CR140	Una aplicación necesita codificar el mensaje y certificar la *Completitud* de los datos.
I8K.COD-CR	Una aplicación necesita codificar el mensaje, evaluar y certificar los mensajes de datos maestros acorde a los niveles de calidad de *Precisión* y *Completitud*.

Tabla 1. Tipos de mensajes de datos maestros intercambiados entre las aplicaciones e I8K

Tipo	Descripción
I8K.REQ	Una aplicación envía un mensaje de petición de datos a un proveedor de datos.
I8K.RES	Un proveedor de datos envía de vuelta un mensaje de respuesta con los datos que ha solicitado una aplicación.

Tabla 2. Tipos de mensajes de datos maestros intercambiados entre aplicaciones

3 Propuesta

El objetivo de este artículo es describir la modificación y ampliación de la arquitectura I8K [10] para dar soporte a la evaluación de grandes volúmenes de datos sin perjuicio de su rendimiento. Para ello, se asume que los datos podrán ser procesados en bloque

– no de forma continua o streaming –. Esta suposición permite utilizar el software Hadoop [12] y la pila tecnológica asociada – como HDFS o el modelo de programación Map-Reduce [13]. En la **Fig. 4** se presentan los cambios realizados a la arquitectura I8K para disminuir el impacto del volumen de los datos en la evaluación de las dimensiones de calidad de datos de *Precisión* (ISO/TS 8000-130) y *Completitud* (ISO/TS 8000-140).

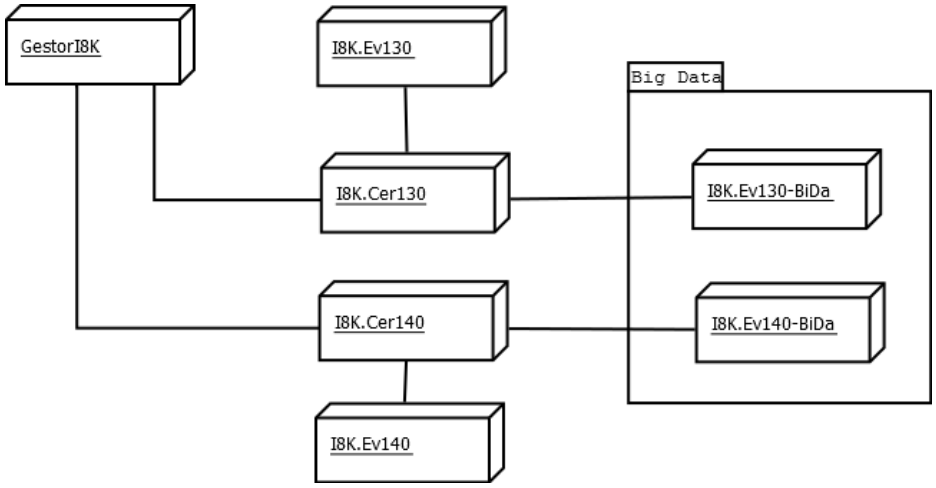

Fig. 4. Arquitectura I8K|DQ-BigData

Nuestra propuesta no elimina la capacidad de la arquitectura para la evaluación de calidad de datos regulares, sino que añade la capacidad de evaluación de grandes volúmenes de datos utilizando tecnologías Big Data. Este hecho, aunque trivial, puede añadir una ligera problemática en el uso de la arquitectura I8K. El conjunto de mensajes mostrados en las tablas 1 y 2, no permiten al GestorI8K conocer qué tipo de datos están incluidos en el Mensaje de Datos Maestros, y por lo tanto, el GestorI8K no sabría dónde enviar los datos para su evaluación (i.e., enviar a I8K.Ev-1x0 o a I8K.Ev-1x0-BiDa). Para garantizar el adecuado funcionamiento de la nueva arquitectura, es necesario añadir nuevos mensajes que permitan especificar a I8K que debe utilizar los servicios nuevos para el caso de Big Data, en lugar los servicios "clásicos" usados en un contexto de datos regulares. Estos nuevos tipos de mensajes están reflejados en **Tabla 3**.

Tipo	Descripción
I8K.CODED-BiDa	I8K ha codificado un mensaje de datos maestros para Big Data y el contenido es devuelto a la aplicación solicitante.
I8K.CR-BiDa	Una aplicación necesita codificar el mensaje, evaluar y certificar los mensajes de datos maestros acorde a los niveles de calidad de *Precisión* y *Completitud* para Big Data

I8K.CR130-BiDa	Una aplicación necesita codificar el mensaje, evaluar y certificar la *Precisión* de los datos para Big Data.
I8K.CR140-BiDa	Una aplicación necesita codificar el mensaje y certificar la *Completitud* de los datos para Big Data.

Tabla 3. Nuevos Tipos de Mensajes

Con estos nuevos mensajes, el protocolo de comunicación no cambia, sino que tan sólo es necesario especificar e incluir dichos mensajes. Para facilitar la comprensión del artículo, se reproduce a continuación el funcionamiento de la arquitectura. El concepto de *"codificación de datos"* se corresponde con la operación de hacer un Mapping de la definición de los Datos Maestros proporcionados por una de las organizaciones a la definición estándar de dato maestro que está en el vocabulario del diccionario de datos de I8K.

1. Una aplicación A solicita a I8K codificar un mensaje para realizar una petición de datos a un proveedor de datos B.
2. I8K codifica el mensaje de A.
3. A envía el mensaje codificado solicitando los datos a B.
4. B pide a I8K que decodifique dicho mensaje.
5. I8K decodifica el mensaje para B.
6. B procesa el mensaje y dependiendo de si los datos son Big Data o no:
   (a) Envía mensaje para procesar los datos de manera regular.
   (b) Envía mensaje para procesar los datos con la extensión Big Data.

4 Ejemplo de aplicación

Una empresa bancaria denominada "*Bancancha*" tiene que intercambiar datos con otra empresa subsidiaria. Ambas empresas han solicitado incluir información correspondiente al nivel de calidad de datos de la Precisión y de la Completitud. Para ello, se considera necesario el uso de I8K. Además, dado que se espera un gran volumen de datos, se considera prioritaria la utilización de nuevos servicios BiDa- I8K para certificar el nivel de calidad que tienen dichos datos.
Para este ejemplo de aplicación, se ilustrarán los resultados obtenidos de aplicar la evaluación 140, Completitud.

La arquitectura fue desplegada en una máquina *host* junto con tres máquinas virtuales con una instalación de Hadoop (se asume que los datos pueden ser procesados con tecnologías Hadoop). Una de las máquinas virtuales hará de máquina maestra y las otras dos tomarán el papel de esclavas.
En la máquina *host* se ejecutan los servicios web de I8K (GestorI8K) que realizan las correspondientes llamadas a los certificadores. Los certificadores a su vez, redirigen los

mensajes hasta los evaluadores para obtener los niveles de Calidad de Datos de los datos incluidos en el Mensaje de Datos Maestros. Estos evaluadores son ejecutados en las máquinas virtuales, y están desarrollados según el modelo de programación Map-Reduce, están implementados en Python (ver Anexo A [16]). Los evaluadores están encapsulados en servicios web ejecutados a través del framework Hadoop Streaming [17]. Los resultados de evaluación son devueltos por el camino descrito hasta el GestorI8K.

El Mensaje de Datos Maestros cuenta con 1056321 registros y en el Anexo A [16], se muestra lo que podría ser un fragmento del mismo. Este mensaje debe contener un apartado en el que se especifiquen las reglas de calidad de datos definidas por las organizaciones implicadas en el intercambio de datos. En el caso de la Completitud, es necesario que se defina en una regla los términos (del diccionario de Datos Maestros) que se consideran necesarios (i.e., es necesario que exista un valor para cada uno de ellos en cada registro).

La operativa de lo que se sigue es la siguiente:

1. La empresa *"Bancancha"* solicita la certificación de Completitud de los datos maestros que se necesitan intercambiar. Para ello se envía a I8K un mensaje de datos maestros del tipo I8K.CR140-BiDa (descrito en la **Tabla 4**).
2. El GestorI8K recibe el mensaje de la empresa proveedora de los datos y, al interpretar el Mensaje de Datos Maestros, decide que tiene que invocar al servicio I8K.Cer140 para certificar el nivel de Completitud de los datos contenidos en el Mensaje de Datos Maestros.
3. El módulo I8K.Cer140 solicita los servicios del Evaluador 140, y al ser de tipo Big Data reenviará el mensaje al servicio I8K.Ev140-BiDa.
4. Cuando I8K.Ev140-BiDa recibe el mensaje, procede a evaluar los datos contenidos en él. En primer lugar almacena los 1056321 registros de datos en la etiqueta <data> del Mensaje de Datos Maestros en el HDFS [18] asociado al despliegue de Hadoop. Para llevar a cabo la evaluación, invoca al servicio web que encapsula el programa con el evaluador de completitud implementado utilizando modelo de programación Map-Reduce.

El módulo *I8K.Ev140-BiDa* se compone de un Mapper y un Reducer. El Mapper comprueba línea a línea los registros en los que existen valores para los términos (campos del registro) necesarios para la organización según la regla definida. Cada registro procesado es marcado indicando si cumple la regla (i.e., tiene todos los valores para los términos necesarios) o no. Además, se añaden los términos que a pesar de no estar en la regla, no están vacíos.

El Reducer recoge los resultados del Mapper y comprueba línea a línea el tipo de término (necesario para la organización o no – i.e., en la regla o no) para poder calcular su nivel de Completitud. Para el cálculo de la Completitud se manejan los siguientes conceptos:

- *Nivel mínimo de Calidad*: Nivel mínimo de completitud/precisión que requiere la organización. Se cumplirá el nivel mínimo sólo si se cumplen todas las reglas, i.e., si todos los términos necesarios para la organización según las reglas definidas, tienen un valor válido asociado.
- *Términos sin regla:* Número total de términos que no aparecen en las reglas (i.e., no son necesarios para la organización), pero que forman parte del mensaje.
- *Términos no vacíos sin regla:* Número *Términos sin regla* que, además, no están vacíos (i.e., tienen un valor válido asociado), pero que forman parte del mensaje.
- *Porcentaje restante*: Porcentaje restante de Completitud obtenido de la operación de restar, al total (100%), el *Nivel mínimo de Calidad*.

Para calcular el nivel de Completitud (lvl140) de cada registro se realiza el siguiente cálculo:

$$Completitud = Nivel\ mínimo\ de\ Calidad + \frac{Porcentaje\ Restante * Términos\ no\ vacios\ \text{sin}\ regla}{Términos\ \text{sin}\ regla}$$

El valor de Calidad, que en este ejemplo es el valor de Completitud, es el que se debe adherir a cada registro de datos en el Mensaje de Datos Maestros, puesto que representa la información sobre la calidad asociada a los Datos Maestros que se van a intercambiar. El mensaje con la información de Calidad de Datos adherida se debe enviar de vuelta al GestorI8K de la arquitectura. La empresa *"Bancancha"* podría entonces decidir qué datos intercambiar, por ejemplo, datos con un nivel de calidad por encima del umbral del 85%.

En un tiempo inferior a 1 minuto, se obtienen los resultados parciales y el total de la Completitud de más de 1 millón de los Datos Maestos procesados. En comparación, sin las tecnologías Big Data, I8K podía tardar entre 25-30 minutos en condiciones similares. Si bien es cierto que estos resultados son positivos desde el punto de vista de validación de la solución, hay que tener en cuenta que el ejemplo se ha practicado en un entorno controlado. Para validar totalmente la arquitectura I8K|Big Data, es necesario un caso de estudio completo en el entorno real de una o varias empresas.

Una vez que se obtienen los resultados, se envían al módulo I8K.Cer140 que prepara un XML con la información de la certificación de Calidad de Datos y de los niveles de Completitud obtenidos en la evaluación.

5 Conclusiones y trabajos futuros

Es innegable que vivimos en una época en donde los datos tienen un valor fundamental. Las organizaciones intercambian datos para ejecutar sus procesos de negocio. Es necesario que los datos tengan niveles adecuados de calidad para que esos procesos puedan ejecutarse con éxito. Este éxito podría asegurarse si se pudieran filtrar los datos por su nivel de calidad, y este nivel de calidad debería estar incluida junto con los mismos datos intercambiados. ISO/TS 8000 partes 100 a 140 pueden ayudar en este cometido. I8K es una arquitectura orientada a servicios que implementa los requisitos de las partes

de I8K anteriormente mencionadas. Esta implementación no ofrece un buen rendimiento cuando tiene que procesar volúmenes grandes de datos. Para paliar estas deficiencias, se han añadido nuevas partes que usan tecnologías de Big Data bajo unos supuestos de partida. Este artículo describe cómo se ha utilizado Hadoop y el modelo de programación Map-Reduce para desarrollar de I8K en entornos Big Data. Estos nuevos evaluadores permiten el análisis de las dimensiones de calidad datos Precisión y Completitud en grandes volúmenes de datos con un rendimiento adecuado.

Los fundamentos obtenidos se han aplicado a un ejemplo, y se ha comprobado la efectividad de los cambios en la arquitectura I8K con resultados positivos en la eficiencia del proceso (de 25-30 minutos a menos de un minuto).

Como trabajos futuros, esta arquitectura puede ser ampliada para la evaluación y certificación de otras dimensiones de calidad de datos. También se propone la mejora de la arquitectura I8K|Big Data, con tecnologías en tiempo real como Apache Storm.
Por último, pero no menos importante, es importante verificar la aplicabilidad de la arquitectura I8K|Big Data en empresas reales, para comprobar el grado de usabilidad de la arquitectura, no sólo en entornos académicos, sino también en el mundo real de las empresas.

6 Reconocimientos

Este trabajo ha sido financiado por el proyecto GEODAS-BC (Ministerio de Economía y Competitividad y Fondo Europeo de Desarrollo Regional FEDER, TIN2012-37493-C03-01); proyecto SERENIDAD (Consejería de Educación, Ciencia y Cultura de la Junta de Comunidades de Castilla La Mancha, y Fondo Europeo de Desarrollo Regional FEDER, PEII-2014-045-P); proyecto VILMA (Consejería de Educación, Ciencia y Cultura de la Junta de Comunidades de Castilla La Mancha, y Fondo Europeo de Desarrollo Regional FEDER, PEII-2014-048-P); proyecto GLOBALIA (Consejería de Educación, Ciencia y Cultura de la Junta de Comunidades de Castilla La Mancha, de la Junta de Comunidades de Castilla La Mancha, y Fondo Europeo de Desarrollo Regional FEDER, PEII-2014-038-P) y CGT – DESARROLLO GLOBAL DEL SOFTWARE (Ref.: CGT140050, Orgánica: 00541R040).

7 Referencias

[1] S. Mohanty, M. Jagadeesh, and H. Srivatsa, *Big Data Imperatives*, 2013.
[2] T. C. Redman and A. Blanton, *Data quality for the information age*: Artech House, Inc., 1997.
[3] D. Loshin, *Master data management*: Morgan Kaufmann, 2010.

[4] ISO/TS, "ISO 8000-100: Data Quality - Part 100: Master data: Exchange of charateristic data: Overview," ed, 2009.
[5] ISO/TS, "ISO 8000-110, Data quality - Part 110: Master data: Exchange of characteristic data: Syntax, semantic encoding, and conformance to data specification.," ed, 2009.
[6] ISO/TS, "ISO/TS 8000 -120, Data quality - Part 120: Master data: Exchange of characteristic data: Provenance," ed, 2009.
[7] ISO/TS, "ISO/TS 8000-130, Data quality — Part 130: Master data: Exchange of characteristic data: Accuracy," ed, 2009.
[8] ISO/TS, "ISO/TS 8000-140, Data quality — Part 140: Master data: Exchange of characteristic data: Completeness," ed, 2009.
[9] I. Caballero, I. Bermejo, L. Parody, M. T. G. López, R. M. Gasca, and M. Piattini, "SLA4DQ-I8K: Acuerdos a Nivel de Servicio para Calidad de Datos en Intercambios de Datos Maestros regulados por ISO 8000-1x0," *JCIS*, 2014.
[10] I. Caballero, I. Bermejo, M. T. G. López, R. M. Gasca, and M. Piattini, "I8K: AN IMPLEMENTATION OF ISO 8000-1X0 " *17th International Conference on Information Quality (ICIQ)*, 2013.
[11] C. P. Chen and C.-Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information Sciences*, vol. 275, pp. 314-347, 2014.
[12] The Apache Software Foundation. (04/05/2015). *Apache Hadoop*. Available: https://hadoop.apache.org
[13] The Apache Software Foundation. (2015). *Map Reduce*. Available: https://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html
[14] A. Borek, A. K. Parlikad, J. Webb, and P. Woodall, *Total information risk management: maximizing the value of data and information assets*: Newnes, 2013.
[15] I. Bermejo, "PROYECTO FIN DE CARRERA I8K: Arquitectura de Servicios para la Gestión de la Calidad de los Datos: Una implementación de ISO 8000:2009-100," 2013.
[16] B. Rivas, J. Merino, I. Caballero, M. Á. Serrano, and M. Piattini. (2015). *Anexo A: Propuesta de Desarrollo del análisis de soporte a la etapa de Evaluación de Calidad de Datos*. Available: http://alarcos.esi.uclm.es/download/jisbd2015-paper73-anexoA.pdf

[17] The Apache Software Foundation. (2015, 02/02/2015). *Hadoop Streaming*. Available: https://hadoop.apache.org/docs/current/api/org/apache/hadoop/streaming/package-summary.html