DISEÑO DE PLATAFORMAS TECNOLÓGICAS PARA ANALÍTICA DE BIG DATA EN SISTEMAS CIBER-FÍSICOS INDUSTRIALES

Eduardo Antonio Hinojosa Palafox[1], Oscar Mario Rodríguez Elías[1], José Antonio Hoyo Montaño[1], Sonia Regina Meneses Mendoza[1]

[1]Tecnológico Nacional de México /I. T. de Hermosillo, División de Estudios de Posgrado e Investigación
e-mail: d02330027@hermosillo.tecnm.mx, omrodriguez@hermosillo.tecnm.mx, jose.hoyom@hermosillo.tecnm.mx, sonia.menesesm@hermosillo.tecnm.mx

***Resumen*— A través de los sistemas de analítica industrial es posible identificar ideas, patrones o modelos útiles necesarios para la innovación sostenible. la creación de plataformas tecnológicas para promover servicios de optimización enfrenta los desafíos de los sistemas ciberfísicos industriales que deben considerar un enfoque novedoso para el diseño de una arquitectura de referencia que integre la convergencia de tecnologías en la analítica de Big Data industrial y el aprendizaje máquina. Este trabajo de investigación presenta un enfoque metodológico para el diseño de una arquitectura de referencia para analítica de Big Data industrial que provee servicios de optimización para la detección temprana de fallas en la industria 4.0 a través de los métodos basados en datos. La arquitectura de referencia fue validada en un escenario de Big Data industrial que incorpora dos servicios basados en almacenamiento HDFS. El primer servicio, Data Analytics Studio (DAS), extrae información basada en consultas SQL que permite generar vistas y nuevas tablas. El segundo servicio permite el análisis con Spark mediante un cuaderno de trabajo Zeppelin basado en la web para el análisis de datos en forma interactiva. Finalmente, se ha definido un marco de trabajo que sirve para agilizar y facilitar el diseño de soluciones de Big Data industrial, con una metodología de diseño para una arquitectura que permita integrar fases y herramientas para brindar soluciones a escenarios de uso concretos.**

Palabras Clave: Industria 4.0, manufactura 4.0, Big Data Industrial, Sistemas ciber-físicos industriales, analítica industrial.

I. INTRODUCCIÓN

La industria está migrando de un enfoque tradicional a uno en el que una máquina no solo se limita a producir, sino que debe hacerlo de una manera inteligente y energéticamente eficiente, también debe ser capaz de proporcionar información sobre el proceso a varios rangos de la jerarquía de la organización (Drath & Horch, 2014).

Este nuevo enfoque conocido como Industria 4.0 marca un hito importante en el desarrollo industrial y expresa la idea de que se está en el comienzo de una cuarta revolución industrial (Zezulka et al., 2016). Su base es que la conexión de máquinas, sistemas y activos en las organizaciones pueden crear redes inteligentes a lo largo de la cadena de valor para controlar los procesos de producción (Ochs & Riemann, 2017). En este nuevo escenario, el foco no está solo en las nuevas tecnologías sino también en cómo se combinan desde la perspectiva de los datos.

Esta oportunidad de interconexión trae consigo la posibilidad de que la generación constante de grandes volúmenes de datos sea utilizada en diversas aplicaciones de la industria, aunque también trae desafíos, como la necesidad de un modelo novedoso que tenga en cuenta cambios en la convergencia de tecnologías en sistemas ciberfísicos industriales (iCPS, por sus siglas en inglés de industrial cyber-physical systems) (Lee et al., 2014), además de considerar la importancia de un enfoque de diseño centrado en el modelo de datos que permita crear una arquitectura de referencia para el desarrollo de la analítica industrial de Big Data (Lee et al., 2015a).

El Internet de las Cosas (IoT, por las siglas en inglés de Internet of Things) está siendo ampliamente incorporado en la industria, y su impacto la está transformando (Lade et al., 2017). El Internet Industrial de las Cosas (IIoT, por las siglas en inglés de Industrial Internet of Things) consiste en crear redes de objetos físicos, entornos, vehículos y máquinas a través de dispositivos electrónicos integrados que permiten la recopilación e intercambio de datos (Madakam, S., Ramaswamy, R. and Tripathi, 2015). El IIoT abre la posibilidad a los iCPS de converger con la tecnología de la información y relacionado a redes, conectividad, datos, ecosistemas y sistemas de información con tecnología operativa hablando de equipos físicos de la planta, maquinaria, sistemas de monitoreo y control. Es decir, la construcción de iCPS permite la integración de la información con los procesos industriales (Yan, J., Meng, Y., Lu, L., & Li, 2017). En este contexto, los iCPS conectan máquinas, sistemas, activos y organizaciones para crear redes inteligentes a lo largo de la cadena de valor, entre otras cosas, para optimizar los procesos de producción (Jeschke et al., 2017).

La computación en la nube de forma general puede entenderse como un enfoque en el uso de recursos informáticos (hardware y/o software), a los que se accede a voluntad mediante la contratación de servicios a terceros (Vora et al., 2016). El enfoque principal de la computación en la nube industrial es la integración y las soluciones verticales en lugar de las horizontales, que es el foco del cómputo en la nube general, esto significa que las soluciones de nube industrial se centran en crear más valor dentro de los límites de la industria en lugar de ampliar sus alcances. Dado que los dispositivos están conectados a una red amplia, requieren un entorno que permita reunirse e interactuar entre sí ofreciendo y requiriendo servicios. El cómputo en la nube facilita el almacenamiento, procesamiento y gestión de Big Data en

iCPS (Givehchi et al., 2013). La integración con IIoT puede llevar el procesamiento de flujos de datos de detección al siguiente nivel para proporcionar servicios de detección omnipresente más allá de las capacidades de las cosas individuales (Huang et al., 2013).

No obstante, este enfoque involucra aspectos importantes a considerar, los cuales se describen a continuación:

- Los proyectos de Big Data industrial, en general, adolecen de un largo tiempo de desarrollo y ejecución de proyectos, lo que hace que muchos proyectos interesantes no sean económicamente atractivos.
- La gestión de datos en el contexto de la industria 4.0 es un tema complejo que requiere una propuesta que considere la gestión de datos como un componente nuclear en el diseño de una arquitectura de referencia para la analítica industrial.
- Esta propuesta debe considerar la convergencia de IIoT, con el cómputo en la nube en la industria y el modelado basado en datos para el diseño de una arquitectura de referencia que sea la base para el desarrollo de soluciones de analítica industrial en el contexto de los iCPS.

Con base en lo anterior, este artículo de investigación presenta un modelo de gestión de datos que provee servicios de analítica en la industria a través de una arquitectura de referencia flexible y adaptable de analítica de Big Data industrial basada en la gestión de modelos de datos para extraer el conocimiento de los datos generados por los sistemas ciberfísicos industriales. También presenta el diseño basado en software de una solución para iCPS para un sistema de diagnóstico de fallas que ilustra la analítica de Big Data Industrial en sistemas iCPS.

El resto de este documento está estructurado de la siguiente manera: la Sección 2 trata los conceptos generales relacionados con el Big Data industrial. La sección 3 se aborda el plateamiento del problema y la propuesta de solución. La Sección 4 presenta una metodología de diseño para la arquitectura de gestión de datos. En la sección 5, se presenta la arquitecturta de referencia para analítica de Big Data industrial y su validación a través de un escenario de diagnóstico de fallas para servicios de datos de optimización. Finalmente, la conclusión de este documento se encuentra en la Sección 6.

II. Big Data en la Industria

En esta subsección se caracteriza el ciclo de vida del Big Data de manufactura considerando los impulsores de cambio para la gestión de datos en la industria 4.0: El internet industrial de las cosas, el computo en la nube para la industria y el Big Data industrial aplicado en el contexto de los sistemas ciberfísicos industriales.

La industria basada en Big Data necesita adquirir datos de fabricación a gran escala en el ciclo de vida del producto (Ver Figura 1), integrar datos de equipos de la planta y datos externos compartir servicios de datos con los usuarios y, finalmente, lograr la interconexión y la interoperabilidad entre el espacio físico y el espacio cibernético.

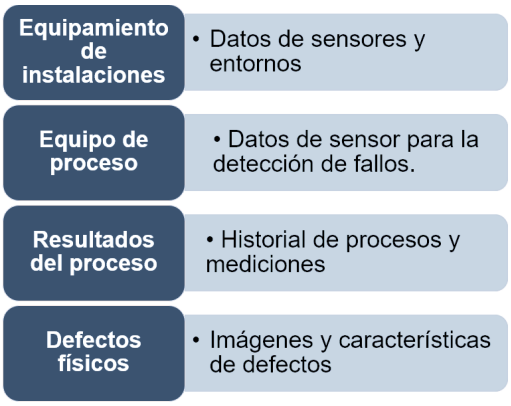

Figura 1. Datos en los procesos de manufactura.

A. *Impulsores de la gestión de Big Data industrial*

Los datos son un activo invaluable en iCPS (Tao et al., 2018), pues permiten la manufactura inteligente. Su importancia estratégica es obtener el valor para la toma de decisiones a través del procesamiento de Big Data.

La acumulación de datos generados en el ciclo de vida de manufactura (Li et al., 2015), planeación, producción y el mantenimiento. Antes de que comience el proceso de producción, el plan de producción inteligente se realiza teniendo en cuenta los datos de recursos del proceso de producción y con fundamento en la relación de los datos globales, el plan de producción global y optimizado puede generarse rápidamente, mejorando la velocidad y precisión de la planeación.

En la fabricación, los datos en tiempo real facilitan el monitoreo del proceso de producción, de modo que los fabricantes puedan mantenerse actualizados sobre las desviaciones de producción para generar planes de control operativo óptimos (Bai et al., 2017). El mantenimiento preventivo activo, a través del almacenamiento y análisis de Big Data del IIoT facilita el diagnóstico de fallas y la optimización del proceso de operación.

A continuación, como se observa en la Figura 2, se describe de forma simplificada el origen de los datos en iCPS para analítica en la industria, que es útil para comprender los impulsores de la gestión de Big Data industrial, incluyendo datos en tiempo real para procesos industriales y datos para sistemas de información de fabricación. En (Tao et al., 2018) se puede encontrar una clasificación completa de los tipos de datos para la fabricación inteligente.

Figura 2. Fuente de datos para analítica industrial.

Datos de recursos del procesamiento de producción, incluidos a) datos recopilados de iCPS por el IIoT; b) los datos de material y producto recopilados de sí mismos y de los sistemas de servicio; c) datos ambientales.

Datos de gestión de los sistemas de información de fabricación (Sistema de ejecución de manufactura (MES, por las siglas en inglés de Manufacturing Execution System), Planificación de recursos empresariales (ERP, por las siglas en inglés de Enterprise Resource Planning), Gestión de relaciones con el cliente (CRM, por las siglas en inglés de Customer Relationship Management), Gestión de la cadena de suministro (SCM, por las siglas in inglés de Supply Chain Software), Método de diagrama de precedencia (PDM, por las siglas en inglés de Precedence Diagramming Method), Sistemas asistidos por computadora (CAS, por las siglas en inglés de Computer Aided Systems), Diseño asistido por computadora (CAD, por las siglas en inglés de Computer-Aided Design), Ingeniería asistida por computadora (CAE, por las siglas en inglés de Computer-Aided Engineering), y la fabricación asistida por computadora (CAM, por las siglas en inglés de computer aided manufacturing).

B. Atributos de calidad del Big Data industrial

Los atributos de calidad en este contexto están relacionados con el diseño de productos de software, con los requisitos funcionales que debe satisfacer el diseño de la arquitectura de software. Por otro lado, los atributos de la calidad de los datos, aunque son una cuestión importante en la calidad de la fabricación, están fuera del alcance de este trabajo y se tratan en otros trabajos, como (Gustavsson & Wänström, 2009).

El esquema propuesto por (Bass et al., 2013), presentado en la Figura 3, describe un atributo de calidad. El estímulo describe un evento que llega al sistema y representa una condición que requiere una respuesta. La fuente del estímulo puede afectar la forma en que el sistema trata el estímulo. La respuesta es la actividad que se realiza en respuesta a la llegada de un estímulo. La medida de la respuesta permite determinar si se cumplió el requisito. El artefacto es la parte del sistema que aplica el requisito. El entorno es el conjunto de circunstancias bajo las que se realiza el estímulo.

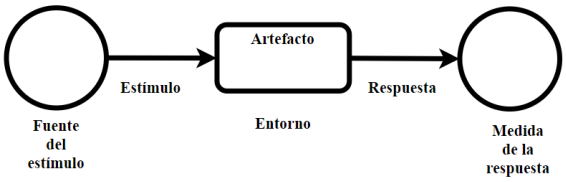

Figura 3. Definición de un escenario.

La técnica de escenarios se centra en identificar el estímulo y cómo el sistema debe responder a él. También se relaciona con los atributos de calidad y busca resaltar las consecuencias de las decisiones arquitectónicas encapsuladas en el diseño. Se consideraron seis escenarios principales para identificar los atributos de calidad que ilustran las características de interés para la gestión de datos en la Industria 4.0. Estos escenarios se propusieron por primera vez en (Hinojosa-Palafox et al., 2019) y se definieron siguiendo la técnica de escenarios para la identificación y descripción de requisitos arquitectónicos de calidad (Kazman et al., 2000).

Con base en lo anterior, los atributos de calidad se definen a continuación:

- Integración de fuentes de datos iCPS. La amplia variedad de sistemas ciberfísicos industriales (iCPS) implementados en una fábrica inteligente generan enormes cantidades de datos. Sin embargo, dado que estos datos provienen de fuentes heterogéneas (PLC, SCADA, ERP), se requiere un sistema ETL para la combinación e integración y almacenamiento posterior en Big Data a gran escala.
- Procesamiento de datos escalable y elástico. Para garantizar el procesamiento de Big Data procedente de tecnologías IIoT (sensores inteligentes, RFID), la arquitectura de nube híbrida tiene que ser escalable y elástica.
- Composición de los eventos basados en datos. Proporciona una estimación de análisis prescriptivo a partir de datos en tiempo real de tecnologías IIoT (sensores inteligentes, RFID) a partir de los parámetros de fabricación esperados en un tiempo de respuesta fiable.
- Servicios de datos de optimización. Proporciona un modelo de análisis predictivo a partir de tecnologías IIoT (sensores inteligentes, RFID) en el procesamiento de Big data en nubes híbridas en un tiempo de respuesta fiable.
- Análisis integrado. Proporcionar algoritmos específicos de análisis de datos adaptados al hardware integrado que genere información cercana al proceso/máquina específica basada en datos generados propios y fuentes de datos en reposo en un tiempo de respuesta confiable.
- Soporte de decisiones basado en análisis. Integración de los datos de fabricación procedentes de las tecnologías IIoT y los Sistemas de Información de Fabricación (MIS) en la toma de decisiones empresariales a través de análisis prescriptivos avanzados, para realizar un análisis paramétrico de los indicadores clave de rendimiento (KPI) del negocio y estimar el error/riesgo o las predicciones de estos KPI.

III. Descripción del Problema y Propuesta de Solución

En la Industria 4.0, los datos son generados por múltiples fuentes en diferentes contextos, como equipos de instalaciones, equipos de proceso, sistemas de fabricación y datos de entorno. Toda esta variedad de datos que llega a alta velocidad y grandes volúmenes se llama Big Data industrial. Como se muestra en la Figura 4, los datos sin procesar son inútiles, por lo que, para obtener la información de los datos, en primer lugar, es necesario limpiar, unificar, consolidar y

normalizar los datos antes de ser procesados debido al ruido, el multiformato, las diferencias de escala, fuentes heterogéneas, entre otros aspectos a considerar en los datos. A continuación, los datos con alto valor se conservan como datos históricos y procesan para el intercambio y el uso compartido en todos los niveles. Por lo general, a través de los servicios en la nube como los servicios de predicción a través de la minería de datos y el aprendizaje automático (Sarnovsky et al., 2018).

Un modelo de gestión de datos para analítica industrial presenta ciertas interrogantes para generar valor:

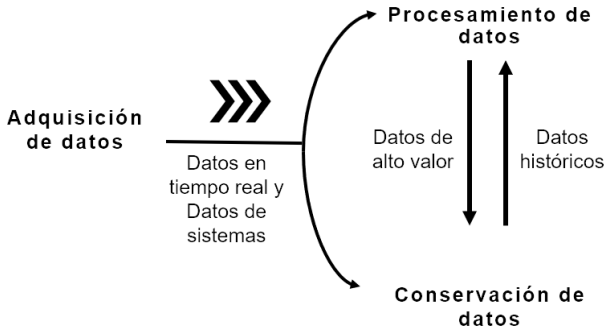

Figura 4. Ciclo de vida del Big Data.

La primera interrogante es ¿cómo abordar los desafíos de los procesos industriales para lograr un modelo de Big Data? relacionada a esta, la segunda interrogante es ¿cómo se lleva la fase de diseño en los sistemas de Big Data industriales? El diseño arquitectónico del sistema de análisis industrial es mucho más crítico que para los sistemas de datos heredados, así que la tercera interrogante se centra en ¿cómo ampliar la arquitectura de diseño tradicional para el diseño del sistema de análisis en iCPS? Una cuarta pregunta considera las anteriores: ¿cómo se pueden integrar el modelo de Big Data y el diseño de arquitectura para el marco del sistema de análisis en el contexto de iCPS?

Diferentes enfoques se pueden encontrar en el diseño arquitectónico para abordar tecnologías específicas o áreas particulares, pero uno que integra una solución cuyo eje central es la gestión de Big Data a lo largo de su ciclo de vida en el contexto iCPS todavía está en desarrollo (Atat et al., 2018; Lee et al., 2015b; O'Donovan et al., 2015). Las soluciones de Big Data para iCPS necesitan integrar tecnologías en un ecosistema consistente en un entorno industrial, pero tales soluciones son complejas. Una guía de arquitectura de referencia para el análisis de Big Data podría facilitar el desarrollo, la implementación y el funcionamiento de las soluciones iCPS de Big Data en la industria (Hinojosa-Palafox et al., 2020).

Se requiere una arquitectura que considere el nuevo paradigma que integra las tecnologías de Big Data en iCPS que permite la creación de arquitecturas a partir de modelos de datos que admitan el análisis de Big Data, para ayudar a abordar los desafíos industriales, considerando que la tecnología de Big Data cambia rápidamente.

La propuesta de solución es el diseño una arquitectura de referencia para el desarrollo de plataformas tecnológicas con apoyo del modelo de gestión de datos desarrollado, siguiendo los requerimientos de analítica para iCPS. La modularidad de la arquitectura permitirá la adaptación a los requerimientos funcionales y no funcionales de los diferentes entornos de iCPS.

IV. Diseño de la Arquitectura de Referencia para Analítica de Big Data Industrial

En esta sección se presenta una metodología basada en el enfoque ADD con la que se diseña una arquitectura de referencia que satisface los requerimientos presentados en la sección II.B (Atributos de calidad del Big Data industrial) y que considere la convergencia del IIoT, el cómputo en la nube y el modelado basado en datos que sirva para facilitar el desarrollo de aplicaciones de analítica industrial en el contexto de los iCPS.

A. Metodología de Diseño de Arquitecturas de software

En el área del diseño de arquitecturas basadas en software, la propuesta de arquitecturas de referencia en diferentes contextos es uno de los enfoques de varios profesionales e investigadores (Jeschke et al., 2017). La determinación de estilos o patrones arquitectónicos, o bien, arquitecturas de referencia, facilitan el diseño de soluciones a problemas que tienen aspectos o características comunes.

Sin embargo, en la revisión del estado del arte, se encontró que las arquitecturas para Big Data en la industria se centran en aspectos relacionados con la gestión de datos aplicados a situaciones específicas. Es decir, en los entornos iCPS, la integración del IIoT con la computación en la nube industrial, incluidos los desafíos de velocidad, volumen, variedad y veracidad, desafían el diseño de sistemas de análisis de Big Data. Por lo que, describir un nuevo proceso de diseño que refleje el cambio necesario para el desarrollo de sistemas de análisis de Big Data que se adapten al cambio de paradigma de la Industria 4.0, es un aporte al estado del arte.

B. El enfoque de diseño basado en atributos

Existen varios métodos de desarrollo de arquitectura de sistemas de software (Capilla et al., 2016), la mayoría de ellos cubren todo el ciclo de vida de la arquitectura y proporcionan pocos detalles sobre cómo realizar la actividad de diseño.

El enfoque de diseño basado en atributos (ADD, por las siglas en inglés de Attribute Driven Design) es el primer método de diseño que se enfoca específicamente en los atributos de calidad a través de la selección de estructuras arquitectónicas y su representación en vistas, también incluye análisis de arquitectura y documentación como parte integral del proceso de diseño (Bass et al., 2002). Las actividades de diseño en ADD pueden incluir refinar los bocetos que se crearon durante las primeras iteraciones de diseño para producir una arquitectura más detallada. ADD comienza con requisitos arquitectónicamente significativos (impulsores y restricciones), y los conecta sistemáticamente con las decisiones de diseño y luego las une a las opciones de implementación disponibles a través de los marcos de referencia (frameworks). ADD también puede usar

arquitecturas de referencia con un catálogo de tecnologías que califique sus atributos de calidad, que incluye tácticas, patrones, entorno de trabajo y tecnologías (herramientas).

V. Resultados Experimentales

En esta sección se presenta la arquitectura de referencia, como se observa en la Figura 5, desarrollada para apoyar el diseño de soluciones de analítica de Big Data industrial. Este aporte original facilita el diseño y la implementación de soluciones basadas en software con un enfoque en gestión de datos para analítica industrial.

Además, se muestra un escenarios de fallas en la industria con datos históricos, para ilustrar un caso de uso en contextos industriales que muestran la aplicabilidad de la arquitectura. También, se revisa una aplicación para detallar las interacciones entre los componentes de la arquitectura propuestos y el caso de uso revisado.

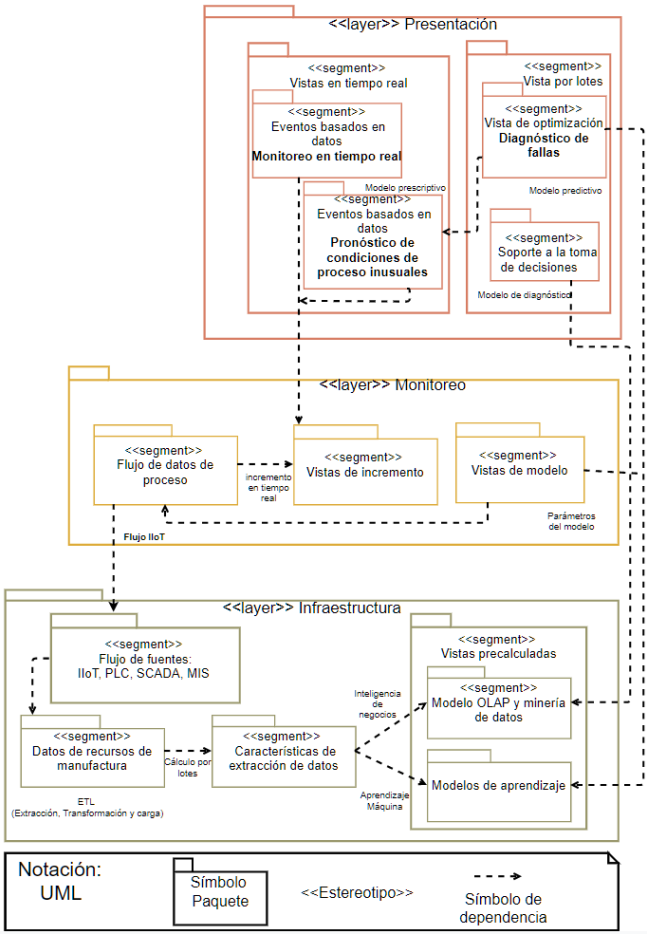
<<layer>> Presentación
<> Vistas en tiempo real
<> Eventos basados en datos Monitoreo en tiempo real
Modelo prescriptivo
<> Eventos basados en datos Pronóstico de condiciones de proceso inusuales
<> Vista por lotes
<> Vista de optimización Diagnóstico de fallas
Modelo predictivo
<> Soporte a la toma de decisiones
Modelo de diagnóstico
<<layer>> Monitoreo
<> Flujo de datos de proceso
incremento en tiempo real
<> Vistas de incremento
<> Vistas de modelo
Parámetros del modelo
Flujo IIoT
<<layer>> Infraestructura
<> Flujo de fuentes: IIoT, PLC, SCADA, MIS
<> Vistas precalculadas
Inteligencia de negocios
<> Modelo OLAP y minería de datos
<> Datos de recursos de manufactura
Cálculo por lotes
<> Características de extracción de datos
Modelos de aprendizaje
Aprendizaje Máquina
ETL (Extracción, Transformación y carga)
Notación: UML
Símbolo Paquete
<<Estereotipo>>
Símbolo de dependencia

Figura 5. Arquitectura de referencia para la gestión de datos para iCPS.

A. *Capa de infraestructura*

La capa de infraestructura es la primera en explicarse. El componente de datos de recursos de fabricación almacena Big Data a partir de una entrada de datos de bajo volumen (por ejemplo, los datos provenientes de bases de datos transaccionales), o datos con alto volumen de entrada (por ejemplo, flujos de datos de sensores), como un conjunto histórico, sólo anexa datos sin procesar. A continuación, se obtiene una lista de características adecuadas para aplicar los algoritmos de aprendizaje de modelos respectivos mediante el componente de procesamiento de datos. Si hay cambios en los criterios, los datos se vuelven a procesar.

En el componente de modelos de aprendizaje, los parámetros que describen el modelo de datos dependen de los métodos de aprendizaje automático utilizados. La detección de anomalías incluiría agregados de sellos de tiempo de la operación de datos de fabricación. En el contexto del procesamiento de grandes conjuntos de datos, es importante utilizar el procesamiento de datos distribuidos en función de los datos reales y el método de aprendizaje utilizado.

El componente modelo OLAP (del inglés Online Analytical Processing: procesamiento analítico en línea) y minería de Big Data permite a través de los almacenes de Big Data el modelado de datos multidimensionales (MDM, por las siglas en inglés de Multi-Dimensional Data Modeling) para el análisis de datos de grandes volúmenes de datos estructurados para apoyar los procesos de toma de decisiones en un contexto de inteligencia empresarial.

B. *Capa de monitoreo*

La capa de monitoreo recibe un nuevo flujo de datos de IIoT, su función principal es analizar el flujo de datos entrante en tiempo real. Se ocupa del procesamiento de datos en tiempo real que suele depender del tiempo, por lo que es crucial reducir las latencias agregadas accediendo al modelo de aprendizaje guardado en la capa de presentación. El componente de flujo de proceso procesa datos de series temporales para obtener las características del nuevo flujo de datos. En el componente de vista de incremento, la medición continua a lo largo del tiempo representa una función importante en la identificación de valores atípicos de datos, se refiere al hecho de que los patrones no se supone que cambien abruptamente, excepto que ocurren procesos inusuales en los datos de trabajo, a continuación, transfiere la entrada del modelo obtenida al componente de monitoreo en tiempo real en la capa de presentación en el período de tiempo coincidente, con el resultado del modelo de aprendizaje y si el umbral es superado, y se detecta consecutivamente durante algún tiempo un evento basado en los datos, se presenta en la vista de tiempo real.

Aunque pueden producirse errores del sensor u otras imprecisiones de datos para difundir un evento de anomalía. Por lo tanto, el evento de anomalía tiene lugar si sucede secuencialmente durante un tiempo específico.

C. *Capa de presentación*

La capa de presentación presenta una vista de la salida de los patrones de datos producidos por las funciones de la capa de infraestructura y de monitoreo a través de vistas en tiempo real y por lotes. Por lo tanto, el componente de monitoreo en

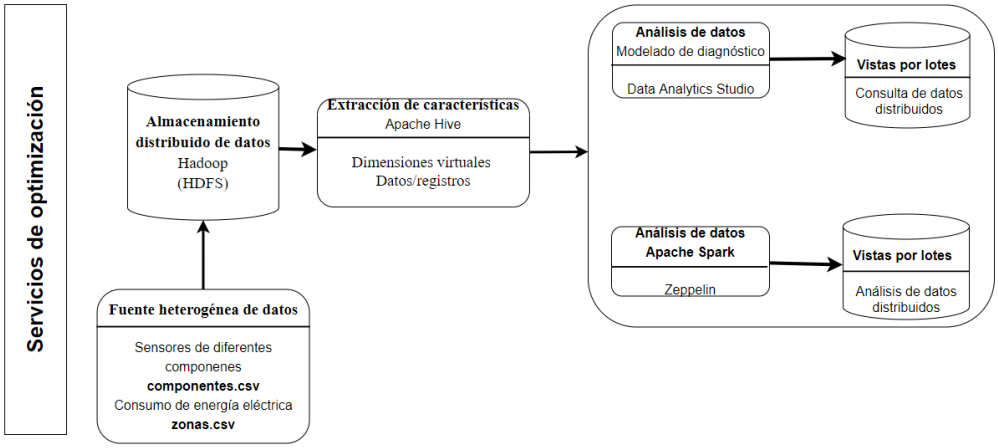

Figura 6. Instancia para analítica de Big Data industrial con datos históricos.

tiempo real analiza la entrega de información actualizada continuamente para identificar anomalías en los datos agregados a lo largo del tiempo de las operaciones de fabricación desde el componente de vistas de incremento, para identificar problemas graves utilizando los parámetros del modelo de aprendizaje del modelo de detección de errores estimado en el componente de vista de modelo. Además, el componente de monitoreo de condiciones de proceso inusuales predice problemas de rendimiento del proceso analizando datos, seguimiento de patrones recurrentes, y detectando comportamientos anormales; utiliza vistas de optimización del modelo de análisis predictivo que se aplica a los eventos basados en datos para el monitoreo proactivo. En las vistas por lotes, el componente de diagnóstico de fallas presenta analítica predictiva para mejorar la detección de anomalías y el diagnóstico de los sistemas de fabricación, con las vistas de optimización de los equipos de proceso y los equipos de instalaciones, basados en métodos de minería de Big Data implementados en el componente de modelos de aprendizaje. El componente de soporte de decisiones basado en análisis utiliza vistas por lotes para admitir el procesamiento analítico en línea (OLAP), análisis para informar, observar y mostrar cuán grande o pequeño es el problema utilizando el almacén de Big Data de la capa de infraestructura.

D. Evaluación de la arquitectura de referencia

La evaluación arquitectónica basada en escenarios es un enfoque bien establecido para validar el diseño arquitectónico y analizar las decisiones que se han tomado para lograr el enfoque de diseño (Raza, Ali; Zafar, Shaista; Rahman, Saeed Ur; Khattak, 2019). Los escenarios son enfoques completos e integrales que reúnen a las partes interesadas de un sistema y las guían a través de un proceso estructurado que explora las opciones de diseño arquitectónico y las implicaciones resultantes.

El propósito de los escenarios es mostrar la sensibilidad de la decisión arquitectónica y los puntos de compensación del diseño arquitectónico. También se describe las interacciones de los componentes de la arquitectura con la aplicación de un caso de uso revisado. Los escenarios brindan una visión del posible ajuste de las soluciones propuestas a los componentes de la arquitectura, considerando hipotéticamente las soluciones como si estuvieran implementadas siguiendo la arquitectura propuesta.

1) Escenario de modelos basados en datos

Las diferentes fuentes de datos pueden dar origen a tres situaciones en el modelado de Big Data y que han sido considerados en el diseño de la arquitectura de referencia: (1) Almacenar los datos en un depósito histórico para su posterior aprovechamiento, (2) Monitorear el flujo de datos y (3) la combinación de datos históricos con el flujo de datos en tiempo real.

En la Figura 6, se muestra una instancia de la arquitectura de referencia propuesta para la analítica de Big Data con datos históricos.

El almacenamiento de Big Data industrial impulsa la necesidad de dividir los datos en distintas computadoras para evitar que una sola máquina se sature. El tipo de sistema de archivos que gestiona el almacenamiento de datos a través de una red de máquinas se llama Sistemas de Archivos Distribuidos. Apache Hadoop está diseñado para almacenar archivos de gran tamaño con acceso a flujo de datos y que se ejecutan en clústeres de hardware básico.

a) Origen de los datos

La sociedad de pronósticos y gestión de la salud (PHM, por sus siglas en inglés Prognostics and Health Management Society) abordó el tema del diagnóstico de fallas de planta industrial con sistemas con datos de registro de fallas incompletos en la Competencia de Desafío de Datos PHM 2015 (Prognostics and Health Management Society, 2015).

Los datos representan: a) series temporales de mediciones de sensores y señales de referencia de control para cada uno de varios componentes de control de la planta (por ejemplo, 6 componentes); (b) datos de series temporales que representen mediciones adicionales de un número fijo de zonas de la planta durante el mismo período de tiempo (por ejemplo, 3 zonas), donde una zona puede abarcar uno o más componentes de la planta; Cada planta se especifica a través de su número de componentes y el número de zonas. La frecuencia de las

mediciones es de aproximadamente una muestra cada 15 minutos, y los datos de la serie temporal abarcan un período de aproximadamente tres a cuatro años.

b) Cargar los datos de sensores en HDFS

Se creó la carpeta sensor-data para almacenar en el sistema de archivos de Hadoop (HDFS) los archivos de datos de una planta industrial: (a) de los sensores de diferentes componentes y sus referencias de control y (b) del consumo de energía eléctrica instantáneo y acumulado.

En la Figura 7, se muestra el sistema de archivos de Hadoop para el escenario de datos históricos de analítica de Big data con datos de sensores industriales.

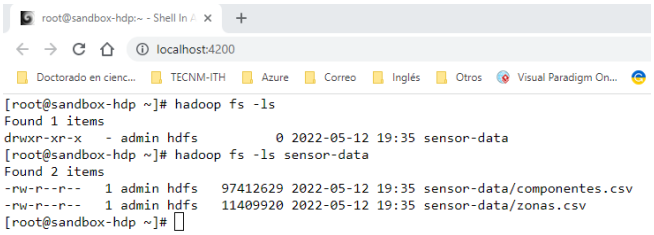

Figura 7. Sistema de Archivos de Hadoop

c) Extracción de características y análisis de datos

Apache Hive facilita el análisis de Big Data industrial almacenado en HDFS con consultas tipo SQL y provee herramientas para la extracción, transformación y carga de datos (ETL, por sus siglas en ingles de Extract, Transform and Load). En la herramienta Data Analytics Studio (DAS) proporciona una interface interactiva para Hive. Los archivos de datos de sensores y mediciones electicas son convertidos en tablas ORC que es un formato optimizado para Big Data y almacenadas en HDFS.

En la Figura 8, se presenta una consulta a la tabla de sensores con una instrucción SQL que calcula el promedio de las observaciones de los cuatro sensores agrupados por fecha y hora.

```
CREATE TABLE sensor_prom
STORED AS ORC
AS
SELECT fechaHora, avg(s1) avgs1, avg(s2) avgs2,
avg(s3) avgs3, avg(s4) avgs4
FROM componentes
GROUP BY fechaHora;
```

Figura 8. Consulta SQL de los datos de sensores en Hive.

Posteriormente los datos son exportados de la tabla ORC a un archivo tipo CSV para el análisis de datos con la herramienta Zeppelin.

d) Análisis de datos con Apache Spark

Zeppelin es un cuaderno de trabajo basado en la web que permite análisis de datos en forma interactiva. Zeppelin aprovecha las características de Apache Spark de extender el modelo de MapReduce, ya que Spark fue diseñado para ser rápida además de una plataforma de cómputo de uso general. Spark tiene integración con Hive, lo que le brinda soporte para los archivos ORC.

e) Usar datos en Hive

El contexto de Hive es una instancia del motor de ejecución SparkSQL que se integra con los datos almacenados en Hive y da soporte SQL en Spark. En la Figura 9 se muestra la creación de un contexto de Hive a través de la variable hiveContext.

```
val hiveContext = new org.
apache.spark.sql.SparkSession.Builder().getOrCreat
e()
```

Figura 9. Consulta SQL de los datos de sensores en Hive.

f) Procesar datos distribuidos

La abstracción del núcleo principal de Spark se denomina conjunto de datos distribuido resistente o RDD (Por sus siglas en inglés Resilient Distributed Dataset). En otras palabras, RDD es una colección inmutable de objetos que se divide y distribuye en varios nodos físicos de un clúster de YARN y que se puede operar en paralelo.

```
/**
La libreria SQL Types nos permite definir los tipos
de nuestro esquema de la base de datos creada en
HIVE y almacenadada en HDFS
/
import org.apache.spark.sql.types._
val sensorPromSchema = new StructType ()
.add("fechaHora", DoubleType, true). add("avgs1",
DoubleType, true
).add("aves2".DoubleTvbe.true).add("aves3".DoubleT
ve.true).add("aves4".DoubleTve.true)
```

Figura 10. Creación de un conjunto de datos distribuidos.

En la Figura 10, se crea un RDD del conjunto de datos almacenados en HDFS desde el archivo CSV que contiene los datos de los promedios de los sensores por fecha y hora, con el esquema que definimos en la tabla.

Ahora es posible poblar el esquema *sensorPromSchema* con los datos del archivo CSV que está almacenado en HDFS, como se muestra en la Figura 11.

```
val sensorPromDataFrame = spark.read.format ("cs")
.option("header", "true") .schema (sensorPromSchema
)load("hdfs:///tmp/data/sensor prom.csv")
```

Figura 11. Poblar el esquema con datos de los promedios de los sensores.

Como se observa en laFigura 12, se crea una vista temporal sensorPromedio.

```
sensorPromDataFrame.createOrReplaceTempView("senso
rPromedio")
```

Figura 12. Vista temporal *sensorPromedio.*

Con lo anterior se usó la sesión de Spark SQL para hacer la consulta a *sensorPromedio,* como se observa en la Figura 13 y realiza la Consulta SQL interactiva a *sensor_prom.*

```
[**
* Inicializa los promedos de los sensores y los
registra como un RDD
*/
val sensor prom = hiveContext.sql("SELECT * FROM
sensorPromedio LIMIT 15")
sensor_prom.createOrReplaceTempView("sensor_prom")
hiveContext.sal("SELECT * FROM sensor prom LIMIT
15").show

Select * from sensor_prom
```

Figura 13. Vista temporal *sensorPromedio* y consulta SQL

En la Figura 14, se muestra el gráfico que permite interactuar a través de cada pestaña que aparecen en la consulta, y ver desplegados un tipo diferente de gráfica dependiendo la configuración de los datos que se deseen.

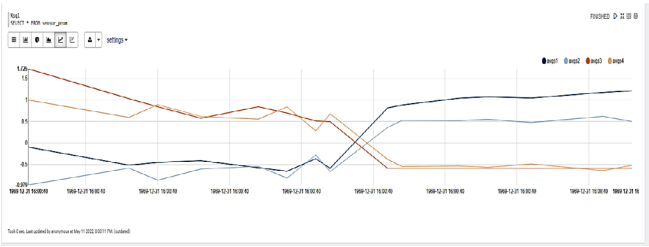

Figura 14. Gráfico interactivo para la consulta SQLen Spark a sensor_prom.

VI. CONCLUSIONES

Este artículo se considera la convergencia tecnológica del Big Data industrial en el contexto de los sistemas ciberfísicos industriales: El Internet Industrial de las Cosas, el Computo en la Nube para la industria y modelado basado en datos. La integración de estos conceptos tecnológicos requiere de nuevos requerimientos de la gestión de datos industrial o atributos de la calidad del Big Data Industrial que debe satisfacer el diseño de una arquitectura de referencia para analítica industrial basada en Big Data. Estos son los requerimientos que cubre el diseño de la mencionada arquitectura de referencia: (1) Integración de fuentes de datos iCPS, (2) Procesamiento de datos escalable y elástico, (3) Composición de los eventos basados en datos, (4) Servicios de datos de optimización, (5) Análisis integrado, (6) Soporte de decisiones basado en análisis.

La metodología para el diseño de una arquitectura de referencia ha sido desarrollada considerando el enfoque de diseño basado en atributos (ADD) y consta de siete etapas: (1) Ciclo de vida del Big Data industrial, (2) Ciclo de vida de fabricación, (3) Impulsores de la gestión industrial de Big Data, (4) Arquitectura en capas, (5) Arquitectura de gestión de datos, (6) Implementación de componentes, (7) Atributos de Big Data industrial.

El uso de este enfoque general permitió el planteamiento de requerimientos funcionales y no funcionales para el soporte del desarrollo de aplicaciones de analítica de Big Data industrial.

Finalmente, para validar la arquitecture de referencia propuesta, se ha implementado el el escenario de datos históricos para analítica de Big Data se ha validado la gestión de los datos generados por los componentes y zonas almacenadas en HDFS. La gestión de los datos masivos de IIoT permite extraer información útil que permite conocer el estado de los sensores de los componentes y el consumo eléctrico de las zonas. Este sistema de analítica de Big Data industrial incorpora dos aplicaciones basadas en almacenamiento HDFS. La primera aplicación, Data Analytics Studio (DAS), extrae información basada en consultas SQL que permite generar vistas y nuevas tablas. De esta manera es posible extraer información almacenada en HDFS de forma estructurada.

Por otro lado, la segunda aplicación permite el análisis con Spark mediante un cuaderno de trabajo Zeppelin basado en la web para el análisis de datos en forma interactiva. De esta forma es posible realizar consultas a la información almacenada en HDFS en forma distribuida. Lo que habilita controlar de los datos para la toma de decisiones.

Además, las plataformas de código abierto utilizadas en la instanciación de la arquitectura de referencia han probado ser adecuadas para el manejo de grandes volúmenes de datos históricos generados por el IIoT. El uso de la API de SparkSQL de Apache Spark para el procesamiento de los datos facilitó una integración más fácil e intuitiva. La abstracción RDD Apache Spark permitió la extracción de un análisis descriptivo de forma rápida.

Con lo anterior, se puede establecer que la arquitectura de referencia ha definido un marco de trabajo que agiliza y facilita el diseño de soluciones para analítica industrial en diferentes escenarios al ser una guía metodológica para integrar fases o etapas con herramientas de software libre que soportan el procesamiento de Big Data de IIoT por lotes y el procesamiento de flujo de datos del IIoT en tiempo real. Por lo que esta implementación muestra que la arquitectura de referencia puede ser utilizada para guiar el diseño de una solución para analítica industrial.

Agradecimientos

CONACYT ha apoyado parcialmente este trabajo con la beca No. 890778, y el Tecnológico Nacional de México bajo la subvención del proyecto 10980.21-P.

REFERENCIAS

Atat, R., Liu, L., Wu, J., Li, G., Ye, C., & Yang, Y. (2018). Big Data Meet Cyber-Physical Systems: A Panoramic Survey. *IEEE Access*, *6*, 73603–73636. https://doi.org/10.1109/ACCESS.2018.2878681

Bai, Y., Sun, Z., Deng, J., Li, L., Long, J., & Li, C. (2017). Manufacturing quality prediction using intelligent learning approaches: A comparative study.

Sustainability, *10*(1), 1–15. https://doi.org/10.3390/su10010085

Bass, L., Clements, P., & Kazman, R. (2013). Software Architecture in Practice, Third Edit. En *Design* (Upper Sadd). Addison Wesley.

Bass, L., Klein, M., & Bachmann, F. (2002). Quality attribute design primitives and the attribute driven design method. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *2290*, 169–186. https://doi.org/10.1007/3-540-47833-7_17

Capilla, R., Jansen, A., Tang, A., Avgeriou, P., & Babar, M. A. (2016). 10 years of software architecture knowledge management: Practice and future. *Journal of Systems and Software*, *116*, 191–205. https://doi.org/10.1016/j.jss.2015.08.054

Drath, R., & Horch, A. (2014). Industrie 4.0: Hit or hype? [Industry Forum]. *IEEE Industrial Electronics Magazine*, *8*(2), 56–58. https://doi.org/10.1109/MIE.2014.2312079

Givehchi, O., Trsek, H., & Jasperneite, J. (2013). Cloud computing for industrial automation systems - A comprehensive overview. *IEEE International Conference on Emerging Technologies and Factory Automation, ETFA*, 1–4. https://doi.org/10.1109/ETFA.2013.6648080

Gustavsson, M., & Wänström, C. (2009). Assessing information quality in manufacturing planning and control processes. *International Journal of Quality and Reliability Management*, *26*(4), 325–340. https://doi.org/10.1108/02656710910950333

Hinojosa-Palafox, E. A., Rodriguez-Elias, O. M., Hoyo-Montano, J. A., & Pacheco-Ramirez, J. H. (2019). Towards an Architectural Design Framework for Data Management in Industry 4.0. *Proceedings - 2019 7th International Conference in Software Engineering Research and Innovation, CONISOFT 2019*, 191–200. https://doi.org/10.1109/CONISOFT.2019.00035

Hinojosa-Palafox, E. A., Rodríguez-Elías, O. M., Hoyo-Montaño, J. A., & Pacheco-Ramírez, J. H. (2020). Trends and Challenges of Data Management in Industry 4.0. En S. X. Zhang J., Dresner M., Zhang R., Hua G. (Ed.), *9Zhang J., Dresner M., Zhang R., Hua G., Shang X. (eds) LISS2019*. Springer Singapore. https://doi.org/https://doi.org/10.1007/978-981-15-5682-1_16

Huang, B., Li, C., Yin, C., & Zhao, X. (2013). Cloud manufacturing service platform for small- and medium-sized enterprises. *International Journal of Advanced Manufacturing Technology*, *65*(9–12), 1261–1272. https://doi.org/10.1007/s00170-012-4255-4

Jeschke, S., Brecher, C., Meisen, T., Özdemir, D., & Eschert, T. (2017). *Industrial Internet of Things and Cyber Manufacturing Systems* (pp. 3–19). https://doi.org/10.1007/978-3-319-42559-7_1

Kazman, R., Klein, M., & Clements, P. (2000). ATAM : Method for Architecture Evaluation. *Cmusei*, *4*(August), 83. https://doi.org/(CMU/SEI-2000-TR-004, ADA382629)

Lade, P., Ghosh, R., & Srinivasan, S. (2017). Manufacturing analytics and industrial Internet of Things. *IEEE Intelligent Systems*, *32*(3), 74–79. https://doi.org/10.1109/MIS.2017.49

Lee, J., Bagheri, B., & Kao, H. A. (2015a). A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems. *Manufacturing Letters*, *3*(October 2017), 18–23. https://doi.org/10.1016/j.mfglet.2014.12.001

Lee, J., Bagheri, B., & Kao, H. A. (2015b). A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems. *Manufacturing Letters*, *3*(December), 18–23. https://doi.org/10.1016/j.mfglet.2014.12.001

Lee, J., Bagheri, B., & Kao, H.-A. (2014). Recent Advances and Trends of Cyber-Physical Systems and Big Data Analytics in Industrial Informatics. *Int. Conference on Industrial Informatics (INDIN)*, *November 2015*, 1–6. https://doi.org/10.13140/2.1.1464.1920

Li, J., Tao, F., Cheng, Y., & Zhao, L. (2015). Big Data in product lifecycle management. *International Journal of Advanced Manufacturing Technology*, *81*(1–4), 667–684. https://doi.org/10.1007/s00170-015-7151-x

Ochs, Th., & Riemann, U. (2017). Smart Manufacturing in the Internet of Things Era. En Cham (Ed.), *In Internet of Things and Big Data Analytics Toward Next-Generation Intelligence* (pp. 199–217). Springer. https://doi.org/10.1007/978-3-319-60435-0_8

O'Donovan, P., Leahy, K., Bruton, K., & O'Sullivan, D. T. J. (2015). Big data in manufacturing: a systematic mapping study. *Journal of Big Data*, *2*(1), 20. https://doi.org/10.1186/s40537-015-0028-x

Prognostics and Health Management Society. (2015). *PHM data challenge 2015*. https://www.phmsociety.org/events/conference/phm/15/data-challenge

Raza, Ali; Zafar, Shaista; Rahman, Saeed Ur; Khattak, U. (2019). Software Architecture Evaluation Methods: A Comparative Study. *International Journal of Computing and Communication Networks*, *1*(2), 1–9.

Sarnovsky, M., Bednar, P., & Smatana, M. (2018). Big Data Processing and Analytics Platform Architecture for Process Industry Factories. *Big Data and Cognitive Computing*, *2*(1), 3. https://doi.org/10.3390/bdcc2010003

Tao, F., Qi, Q., Liu, A., & Kusiak, A. (2018). Data-driven smart manufacturing. *Journal of Manufacturing Systems*, *48*, 157–169. https://doi.org/10.1016/j.jmsy.2018.01.006

Vora, R., Garala, K., & Raval, P. (2016). An era of big data on cloud computing services as utility: 360° of review, challenges and unsolved exploration problems. *Smart Innovation, Systems and Technologies*, *51*, 563–574. https://doi.org/10.1007/978-3-319-30927-9_57

Zezulka, F., Marcon, P., Vesely, I., & Sajdl, O. (2016). Industry 4.0 – An Introduction in the phenomenon. *FAC-PapersOnLine*, *49*(25), 8–12. https://doi.org/10.1016/j.ifacol.2016.12.002