IMF Business School

Hoja de respuestas

Módulo	Módulo II: Impacto y valor del Big Data
Nombre y apellidos	Richard Douglas Grijalba
Fecha entrega	29 Agosto 2023

Indice

Presentación de los casos

1. Diseño de arquitectura para visión cliente 360 º

Diseñar la arquitectura funcional de una plataforma *big data* que proporcione una visión 360 º de los clientes.

En ella, deberían reflejarse:

1-Las fuentes de datos que se van a procesar como entrada. Explicar cuáles de estas fuentes son estructuradas, semiestructuradas y no estructuradas.
2-Diferenciar todas las capas de almacenamiento que tendrá la solución. Explicar y justificar el porqué de cada una de ellas.

2. Analizar el índice de madurez del modelo de negocio de big data en el caso de John Deere

John Deere apuesta la granja a la IA y el IoT

En los últimos 100 años, John Deere, con sus icónicos tractores verdes, ha sido un serio inversor en tecnología.

Comenzó en 1918, cuando Deere compró Waterloo Gasoline Engine Company, y transformó la empresa de un negocio de instrumentos agrícolas en una empresa de tractores industriales de pleno derecho. A finales de la década de 1990, Deere invirtió en GPS con la adquisición de Navcom, que ayudó a allanar el camino para instalar módems 4G LTE en todos sus equipos.

Ahora, cuando llevamos dos décadas de siglo XXI, se trata de inteligencia artificial, aprendizaje automático, redes neuronales, 5G e Internet de las cosas (IoT). En 2017, Deere pagó más de 300 millones de dólares por Blue River Technology, una *startup* de Silicon Valley que está aplicando estas tecnologías al negocio agrícola.

Como vicepresidente senior de Intelligent Solutions Group de Deere, John Stone no solo supervisa la adquisición de Blue River, sino también las otras innovaciones tecnológicas de la compañía, incluida una mayor confianza en la infraestructura como servicio (IaaS) para satisfacer sus necesidades de computación en la nube.

"Puedes pensar en ISG como la división de alta tecnología de John Deere, por lo que tenemos la responsabilidad de desarrollar, diseñar e integrar estas nuevas tecnologías en todos nuestros equipos", dijo Stone a Enterprise Cloud News durante la reciente exposición del Mobile World Congress en Barcelona, donde se interesó por los últimos desarrollos en 5G e IoT.

"Cuando dices IoT, normalmente piensas en cosas que caben en tu bolsillo", dijo Stone, quien ha trabajado para la compañía durante 16 años y tiene experiencia en ingeniería mecánica, así como en programación. "La “T” para nosotros son los tractores de diez toneladas. Nuestro equipamiento ahora tiene módems 4G LTE, con WiFi y Bluetooth, y eso permite una comunicación bidireccional para recolectar datos de la granja y enviarlos a la nube. También toman instrucciones de Deere o de distribuidores u otras compañías de software y las envían a la máquina. La comunicación

bidireccional le dice a la máquina qué hacer. Además, las máquinas también pueden comunicarse entre sí en el campo ".

Se espera que el acuerdo con Blue River lleve esta aproximación a las máquinas industriales y la agricultura al siguiente nivel.

Ahora que el acuerdo entre Deere y Blue River está cerrado, las dos compañías están comenzando a colaborar más, aunque Stone señala que Blue River continuará operando como una empresa independiente.

El futuro de la agricultura

Blue River está desarrollando una tecnología basada en el aprendizaje automático llamada *see and spray*, que puede reducir la cantidad de herbicida utilizado en la agricultura. Específicamente, un rociador toma fotografías de plantas y, mediante el aprendizaje automático y los algoritmos, puede determinar cuáles son malezas y cuáles son cultivos, y solo rocía herbicida sobre las malezas. Eso es importante cuando un campo típico puede tener entre uno y dos millones de plantas.

Este enfoque de la agricultura ofrece una solución doble. La primera es reducir o eliminar la cantidad de herbicida rociado en los alimentos que comen las personas. La segunda es la reducción de costos. Stone estima que un agricultor de soja o algodón podría ahorrar entre 50 y 80 dólares por acre y reducir el gasto en herbicidas hasta en un 90 %.

Esto es solo el comienzo.

"Nuestra hoja de ruta requiere que el aprendizaje automático y la inteligencia artificial se incorporen en cada pieza de equipamiento de John Deere con el tiempo", dijo Stone. "Lo que hacemos con nuestros ojos se puede hacer con mayor precisión con una cámara y una computadora, con un sistema que recuerda esos datos, nunca olvida y se vuelve más inteligente cada vez que pasa por el campo. Esto también se aplica a nuestras divisiones de construcción y equipo pesado".

Además de IA, aprendizaje automático e IoT, hay inversiones en automatización. El ejemplo más obvio son las máquinas autónomas. Sin embargo, también hay otros avances. A finales de año, Deere planea lanzar una aplicación que ofrecerá ajustes de configuración de la máquina desde una ubicación central, y el operador puede aceptar o rechazar esas sugerencias.

A medida que todo esto se va automatizando, una ubicación central puede operar múltiples máquinas en el campo, hacer sugerencias y transmitir información adicional.

Todo esto está respaldado por la nube, específicamente por Amazon Web Services. Además de utilizar la plataforma IaaS de Amazon, Deere es un gran usuario de *Lambda*, la versión AWS de computación sin servidor que permite que las aplicaciones en la nube respondan a diferentes eventos, por lo que utiliza recursos solo cuando es necesario.

El futuro tecnológico de Deere

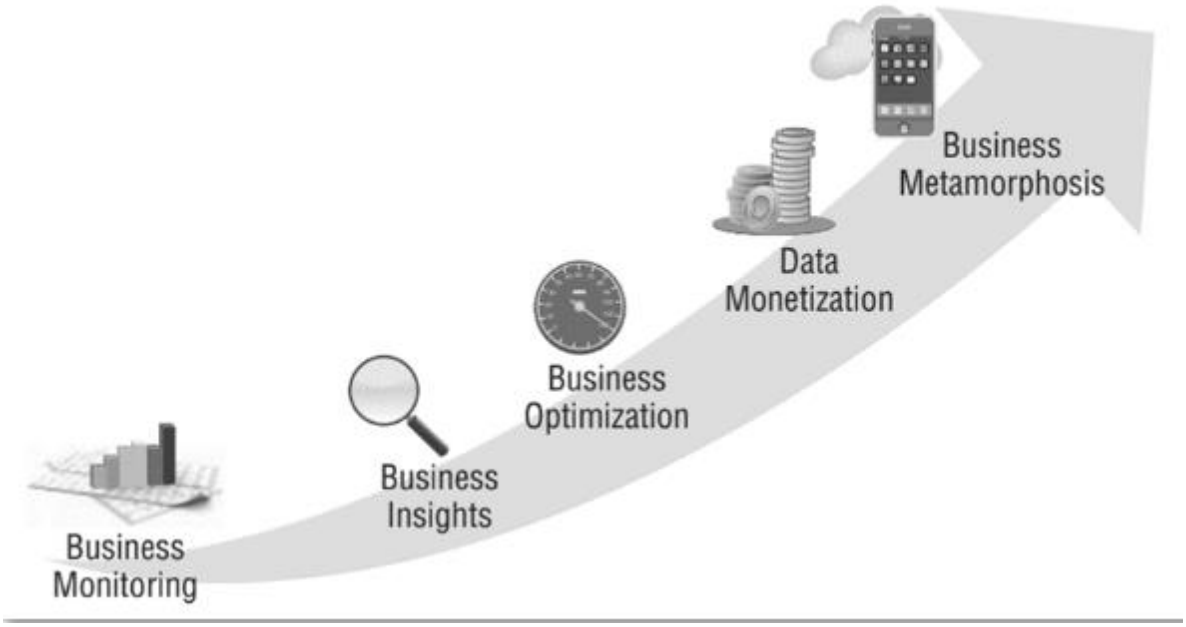
El próximo gran paso para Stone y su equipo sería la introducción generalizada de 5G para acelerar más procesos y disponer de un mayor ancho de banda para admitir aplicaciones más sofisticadas y recopilar mayores volúmenes de datos, por ejemplo, videos de alta definición, ya sea de las máquinas o del mismo campo.

Por ahora, cuando todavía falta un tiempo para un despliegue completo de 5G, Deere se conforma con los módems 4G LTE conectados a cada vehículo, lo que permite usar *edge computing* y descargar datos a la nube.

Como en el pasado, Deere no limita a estos desarrollos tecnológicos sólo para sí, aunque sigue siendo el usuario principal. Por el contrario, es probable que estos desarrollos se comercialicen y vendan a otras compañías, lo que convierte a Deere en un proveedor de tecnología en toda regla.

"Sabemos con certeza que la tecnología es el futuro y que la inteligencia artificial y el aprendizaje automático son el futuro de la agricultura", dijo Stone. "De esta manera, va a ser más eficiente y mejor de lo que es hoy. Dicho esto, continuamos fabricando estas máquinas muy grandes y muy sofisticadas. Mi grupo es sin duda una compañía 100 % tecnológica dentro de Deere".

Leer el artículo y, utilizando el índice de madurez del modelo de negocio de *big data* de la figura, hacer un análisis explicando y justificando en qué etapa estaría John Deere en el momento que se describe en el artículo y qué etapas crees que ha seguido la empresa hasta llegar ahí o, en su caso, qué etapas seguirá posteriormente para completar las etapas descritas en el modelo.

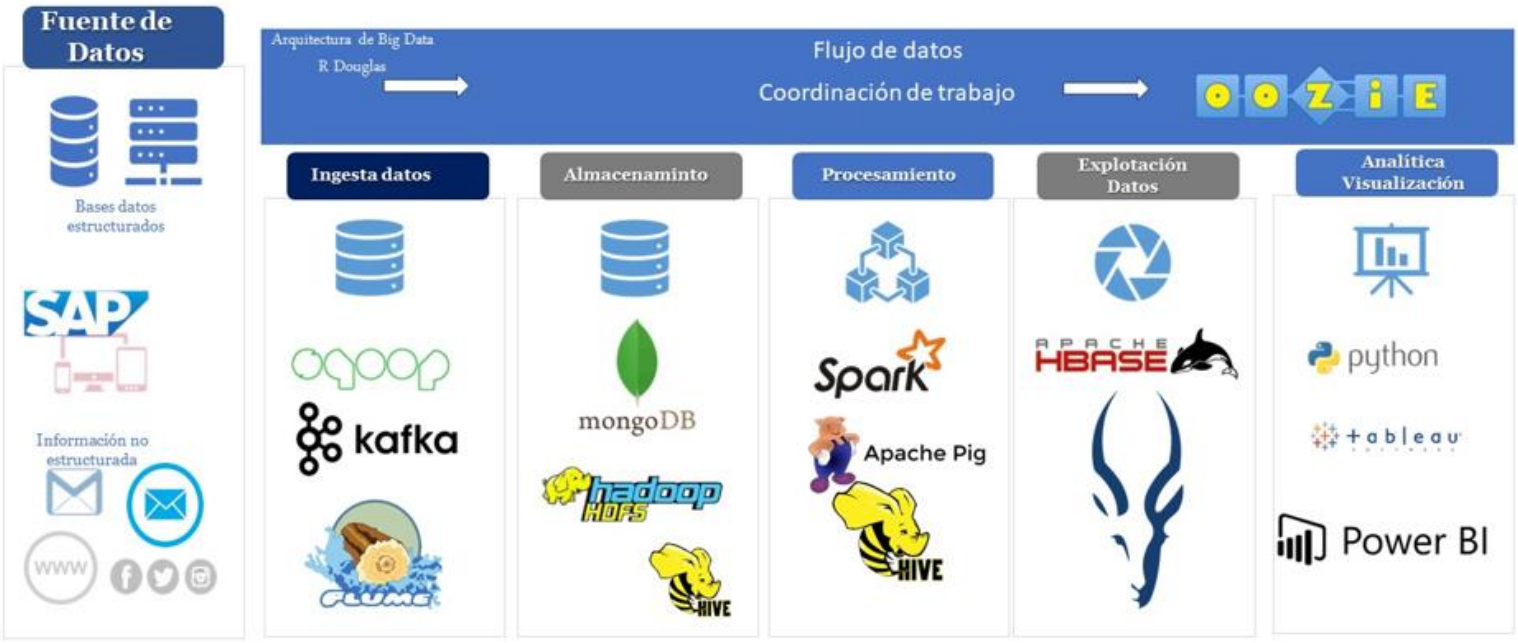


Desarrollo de los casos

Diseño de arquitectura para visión cliente 360 º

La arquitectura primordialmente está basada en las tecnologías más comunes o de conocimientos del mercado, la primordial la plataforma Hadoop, además se agregan algunas que podrían ser de utilizada tales como Mongo DB, Impala, Python, Power Bi, entre otros, en el mercado existen soluciones en la nube (AWS o Google Cloud)o de arquitectura física (granja de servidores), arquitectura pensada en permitir la escalabilidad de los recursos, trabajar en lotes y streaming así como trabajar con datos estructurados y no estructurados, es importante mencionar que también existe la opción Microsoft Azure, la cual presenta una solución para el esquema de Big Data que es interesante al permite enlazar con toda la gama de recursos de la familia Microsoft, para la coordinación el trabajo en este caso interviene oozie.

La arquitectura que se presenta a continuación no pretende ser limitante para ampliar en un futuro los recursos o tecnologías que la empresa que lo requiera, por lo que corresponde a ser una representación de los componentes necesarios en la arquitectura de Big Data, que debe además adoptar una cultura de la *Data*, que le permita ser de esta forma más competitiva en el mercado que se desenvuelva, por lo que para mejorar la funcionalidad de la estructura propuesta, es necesario que la empresa realice una evaluación de la etapa de madurez de cultura y adopción de las tecnologías de Big Data, lo cual colabora al ambiente de control y gobernanza de los datos.

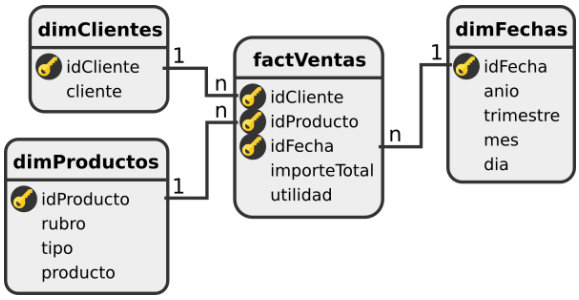


Del recuadro anterior se detalla la propuesta de una estructura de la arquitectura necesaria para cubrir las necesidades de Big Data:

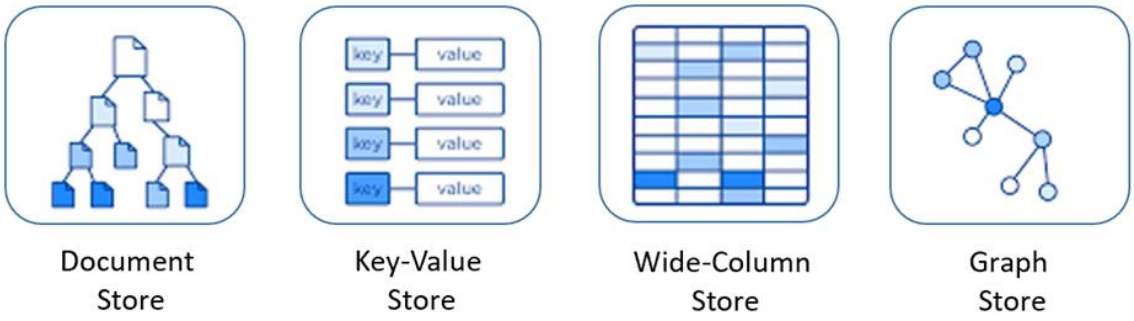
1. Fuente de datos

En atención de las necesidades actuales y futuras que puede tener en requerimientos de datos, la arquitectura de Big Data con visión 360º se establece como parámetro que se trabajarán con datos estructurados y no estructurados.

Datos Estructurados: Datos que presentan una estructura necesaria para ser manejados en bases de datos del tipo SQL, que consisten en tablas con filas y columnas, son un ejemplo perfecto de datos estructurados. Estos datos no pueden ser repetidos en ciertas tablas de dimensión, mientras que las tablas de hechos registran las actividades, se manejan los datos según el modelado estrella.



Datos No Estructurados: Por otra parte los datos no estructurados, no siguen un patrón y provienen de diversas fuentes y de varios tipos de datos, esta es principalmente de una de las características que impulsan el uso de las tecnologías de Big Data, las fuentes de estos datos pueden ser correos electrónicos, carpetas con una serie de archivos variados como imágenes, videos, texto, información proveniente de la web (web scraping) o redes sociales.



2. Ingesta de datos

Una vez que se tiene claro el tipo de datos que se van a recolectar, se procede a definir las tecnologías necesarias para cumplir, este paso denominado **Ingesta de datos** es un componente clave de la arquitectura de big data.

Importante señalar, tal y como se indicó en este módulo *Impacto y valor del Big Data*, en el proceso de ingesta de los datos, **el** uso de **Sqoop** es una buena opción, en vista que es un software que permite extraer información de bases de datos estructuradas e incorporarlas al ecosistema Hadoop.

Kafka interviene como una solución, la cual es una plataforma de streaming de eventos que sirve para **recolectar**, procesamiento y finalmente almacenar datos de eventos de streaming o datos sin principio ni final concretos.

Flume, es un sistema distribuido, confiable y disponible para recoger, agregar y mover grandes cantidades de datos en logs desde diferentes orígenes a un almacén centralizado

3. Almacenamiento

En el almacenamiento se tomó en consideración la importancia de tener la información según los criterios, de disponibilidad y consistencia, tal y como fue señalado en este módulo *Impacto y valor del Big Data,* en el que se indica que, **se debe garantizar la consistencia**, esto implica que los datos repartidos a lo largo de todos los nodos pueden no estar sincronizados. Diferentes usuarios podrían llegar a ver diferentes versiones de un dato.

Garantizar la disponibilidad de la información, esto significa que, en caso de particionado debido a un fallo en la red, el sistema podría quedarse sin responder para evitar así dar una respuesta que no sea consistente.

Con la propuesta se busca utilizar los beneficios del motor de bases de datos como Mongo DB, la cual brinda una posibilidad de dar consistencia en la información y tolerancia al particionado de los datos. Mongo DB es una de las bases de datos NoSQL preferidas por los desarrolladores y profesionales en al área de Big Data.

Hadoop HDFS , Hadoop es un framework opensource para almacenar datos y ejecutar aplicaciones en clusters de hardware básicos, el mismo permite además permite escalar un solo clúster de Apache Hadoop a cientos (e incluso miles) de nodos.

HIVE, es un sistema de almacenamiento de datos para Apache Hadoop. Hive hace posibles el resumen de los datos, las consultas y el análisis de datos. Las consultas de Hive se escriben en HiveQL, que es un lenguaje de consulta similar a SQL

4. Procesamiento de datos

Este proceso es de gran importancia para el resto de la arquitectura de Big Data, en vista que se procesan los datos se realizan transformaciones que permitan a futuro **extraer** información útil para la toma de decisiones, según lo visto durante este módulo, el uso de HIVE, SPARK y PIG, son una serie de herramientas útiles para este desarrollo, esto debido a a que Hive, permite realizar consultas en lenguaje SQL; Pig, que permite escribir rápidamente programas que trabajen con flujos de datos; y Spark,que permite llevar a cabo procesamiento de datos de manera muy rápida.

Procesamiento por Lotes
Se denomina procesamiento por lotes (o procesamiento batch) a la ejecución de un programa que se lanza de manera planificada, sin la supervisión directa de los usuarios.

Procesamiento Streaming
Los datos se van procesando a medida que llegan, o el procesamiento en streaming, donde los datos van llegando en un flujo continuo ininterrumpido.

SparK: Ventajas que nos brinda Spark son muchas, entre ellas la velocidad que permite trabajar los datos e información según como van llegando, en lo que le aventaja a su contraparte HADOOP.

Apache Pig: permite describir el flujo de datos desde entrada sin formato, a través de una o varias transformaciones, para producir el resultado deseado

5. Explotación de datos

Ninguna arquitectura de Big Data es satisfactoria si los datos no se pueden trasladar a los departamentos o procesos que se requieren, por lo tanto la etapa de Explotación de los Datos, es de un carácter necesario y a considerar en toda la gama de herramientas del ecosistema de Big Data.

Esta capa debe incluir una gama de servicios que cubra las necesidades de explotación de la empresa y proporcione acceso a la información a los diferentes usuarios, herramientas y aplicaciones.

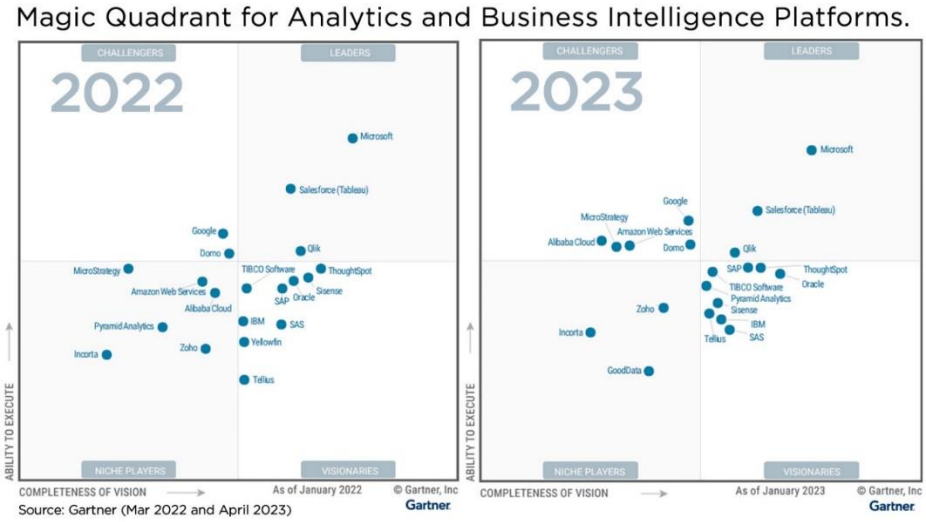
Según el análisis realizado, según el tipo de datos y necesidades, la soluciones que pueden cumplir de una buena forma corresponden a HBASE e Impala, mientras HBASE puede trabajar sin ningún problema con datos NoSQL, Impala nos brinda las soluciones trabajando en lenguaje SQL.

6. Presentación o visualización de datos

En lo que respecta al uso de herramientas para la realización de modelos predictivos, dashboards o reportes, en este caso intervienen Power Bi, Tableau, Python, los cuales tienen características de autoservicio y permiten crear ambientes en el que los usuarios compartas sus resultados con sus grupos de trabajo, sin embargo algunos analistas o científicos de datos preferirán otras herramientas tales como R, el cual es un lenguaje de programación con mucha trayectoria y el cual permite el uso de grandes cantidades de datos.

Cada una de las soluciones y herramientas para visualización de datos pueden ser utilizadas para procesos en menor tamaño de ETL, esto debido a que los datos en bruto o con transformaciones previas podrían no tener la utilidad para el analsita de datos o científico de datos, por lo que Python; R, Power Bi y Tableau permiten realizar extracción, transformaciones y carga de los datos, así como la generación de gráficas y visualizaciones que permiten resumir información en KPI's.

Según el cuadrante de Gartner sobre las tecnologias de Inteligencia de Negocios, Microsoft (Power Bi) lidera el mercado, apareceiendo tambien Tableau, motivos que muestran la necesidad de incorporar ambas opciones.



7. Flujo y Coordinación de Trabajo

Los procesos y comunicación en las diferentes soluciones es algo muy necesario, y esto se logra por medio de una de las soluciones como OOZIE, la cual permite el manejo de flujos de trabajo en el ecosistema de APACHE HADOOP. Esto se traduce como la posibilidad de llevar a cabo proyectos por los desarrolladores en diferentes plataformas, como Map Reduce, Hive o Apache Pig.

Resumén de las etapas o componentes de la arquitectura de Big Data y las posibles tecnologías que pueden intervenir durante cada proceso:

Proceso	Detalle de la actividad	Tecnologías
Ingesta de datos	Variedad de los datos, hace que la recopilación de datos de diversas fuentes y preparación para su almacenamiento y procesamiento sea relevante, los datos pueden ser estructurados o no estructurados, por lo que según el tipo de dato se requiere una solución que cumpla los requerimientos necesarios.	Sqopp, Apache Kafka, Apache Nifi, Apache Flume, Logstash
Almacenamiento de datos	El volumen de los datos, estamos manejan grandes cantidades, el almacenamiento de grandes volúmenes de datos en varios formatos, debe ser de tal forma que se puedan acceder de forma oportuna para su procesamiento y análisis	Hadoop HDFS, Amazon S3, Google Cloud Storage, Mongo DB
Procesamiento de datos	Procesamiento de conjuntos de datos grandes utilizando trabajos por lotes para filtrar, agregar y preparar los datos para su análisis	Apache Spark, Apache Flink, Apache MapReduce, Apache Storm
Explotación de Datos	Los datos están disponibles para la exploración de estos por medio de comandos o herramientas SQL y NoSQL para obtener información valiosa	Apache Hive, Apache HBase, Apache Cassandra, MongoDB
Presentación de datos	Una vez cargado, almacenado, tratado y analizado el conjunto de datos, los mismos están disponibles para los analistas en modalidades de autoservicio, y se llega a utilizar herramientas o paquetes para la Presentación de los resultados del análisis y cuadros de mando Dashboards.	Power BI, Tableau. Python, R
Flujo & Coordinación Trabajo	La coordinación del proceso, flujo correcto, disponibilidad de la información, control de versiones, los mismos se tratan de forma por lotes o por streaming.	Oozie

Analizar el índice de madurez del modelo de negocio de big data en el caso de John Deere

Leer el artículo y, utilizando el índice de madurez del modelo de negocio de *big data* de la figura, hacer un análisis explicando y justificando en qué etapa estaría John Deere en el momento que se describe en el artículo y qué etapas crees que ha seguido la empresa hasta llegar ahí o, en su caso, qué etapas seguirá posteriormente para completar las etapas descritas en el modelo.

Hoy en día se habla cada vez más y más del valor de la tona de decisiones y conocer a las necesidades de los clientes, y se incorpora el concepto de que los Datos son el nuevo petróleo, tal y como lo dijo Clive Humby en el año 2006 “data is the new oil”, en lo que se refiere es que los datos por sí solos no nos son útiles si no se les aplica un procesamiento y refinado (ETL) y posteriormente se generan nuevos productos o ingresos (derivados), es por lo tanto que tal y como lo muestra John Deere a dado ese valor a la información e implementación tecnológica en sus productos, a continuación el análisis de cada una de las etapas de madurez que se describen en el artículo suministrado.

Business Monitoring

Esta esta nos muestra una empresa que comienza a tomar una conciencia sobre el valor de los aportes tecnológicos en el crecimiento y desarrollo de la empresa, John Deere para la época de 90’s, realizó una gran inversión en la tecnología del GPS con la adquisición de Navcom, que ayudó a allanar el camino para instalar módems 4G LTE en todos sus equipos. Este tipo de inversiones se pueden considerar como una visión hacia el future, lo cual algunas grandes compañías en ocasiones carecen de esta visión, sin embargo, este tipo de movimientos pueden hacer la diferencia entre seguir en el mercado o ser desplazado por la competencia, los tractores de John Deere pararon de ser solamente eso *tractores*, sino que pasaron a ser máquinas que generan información en el momento sobre condiciones del funcionamiento de cada uno de los equipos así como de las cosechas, por medio de la incorporación de GPS, Bluetooth y wifi, hoy en día un agricultor o grajero, no sólo puede confiar en la calidad de los tractores sino también en lo útil y facilidades adicionales que los nuevos tractores le brindan, dando esto paso a la siguiente etapa.

Business Insigths

Los negocios generando y recibiendo información, la compañía hace fuertes inversiones, ya sea por medio de compra de soluciones o adquisición de empresas para agregarlas a sus unidades de negocio para el desarrollo de soluciones, la meta en esta etapa corresponde a tener acceso y preparación de datos, tener a disposición las tecnologías que permitan su análisis, y que esto permita descubrir recursos o información para la aplicación de nuevas estrategias, la visión de John Deere de dar el paso en los procesos de transformación digital no es algo que se logra de un año a otro, según el artículo indica es un camino de los último 100 años de existencia de la compañía, Investigación y Desarrollo están plasmados en el la cultura de la empresa, 1990 – 2017, pasar de incorporarle GPS a tener tecnología Machine Learning, posteriormente están implementando tencologías que permiten reducir el uso de plaguicidas por medio de la incorporación de IA & ML, lo cúal hace que se distinga entre una planta comestible y la maleza.

Business Optimization

John Deere, según este artículo previamente observado es el ejemplo de lo que una empresa en nuestros días debería hacer, incorporar tecnología de tal forma que se vuelva en una ventaja competitiva, en vista que tal puede imitar tus habilidades, copiar algo de tu producto, pero no será fácil copiar los Insigths y procesos que rigurosamente se han establecido con el pasar del tiempo, esa cultura del dato, sumado a modelos de Machine Learning para generar análisis prescriptivo, es dado que la compañía invirtió en el año 2017 alrededor de $300 millones en una Startup que para que le suministrara soluciones y tecnologías del tipo de inteligencia artificial, aprendizaje automático, redes neuronales, 5G e Internet de las cosas (IoT).

Data Monetization

El Desarrollo de las plataformas y soluciones que inician para la propia casa, una vez que se han probado y generado un buen avance, se pueden convertir en una nueva fuente de ingresos, por así citarlo un ejemplo de esto sería Amazon que inicialmente sólo vendía libros, pasó a ser la plataforma de mercado virtual y finalmente brinda soluciones en la nube (AWS), de igual forma John Deere, Por ahora, en el cual indica que aún falta un tiempo para un despliegue completo de 5G, Deere se conforma con los módems 4G LTE conectados a cada vehículo, lo que permite usar edge computing y descargar datos a la nube, y esta tecnología se podría conformar en un nuevo ingreso en vista que Deere no estima limitarse a brindar soluciones y tecnologías sólo para sus equipos sino también a futuro espera que estos desarrollos se comercialicen y vendan a otras compañías, lo que convertiría a John Deere en un proveedor de tecnología y soluciones en general.

Business Metamorphosis

Posiblemente haya poco que se pueda describir como una ruta a evaluar o a describir como el futuro, porque John Deere es uno de los que está apuntando en el liderazgo en su mercado, tal parece que John Deere está marcando la pauta de lo que está de moda en el mercado de los equipos verdes, sin embargo si algo se puede apuntar es que el avance que hoy puede ser sorprendente mañana dentro de unos cuantos meses podría parecer descontinuado, por lo que en esta etapa ed *Business Metamorphosis*, John Deere debería apuntar a eso a seguir la hoja de ruta de la innovación y desarrollo, buscar en el mercado la tecnología que se puede o podría aplicar en sus tractores y si las soluciones actuales no satisfacen adecuadamente sus expectativas entonces crear e invertir en el desarrollo de algo novedoso es lo que debe.

Recursos utilizados

1. Material suministrado por IMF, visto en el desarrollo el módulo II Impacto y valor del Big Data
2. Ebook Profundizando en Data Lakes 2019 Morris & Opazo
3. Los logos e imágenes utilizados se obtienen de la página freepng.com y/o encontradas en la web con el fin de ilustrar los temas.
4. Página Web Microsoft learn: Arquitecturas de Big Data

https://learn.microsoft.com/en-us/azure/architecture/data-guide/big-