Asignatura	Datos del alumno	Fecha
Seguridad en Bases de Datos y Almacenamiento de Datos Masivos	GRUPO 29	30.07.2022
	Braulio David Velasco Castillo Blanca Paola Toledo Martínez Angel Ramon Paz López	

Actividades

Trabajo Grupal: Diseño de una arquitectura con seguridad en un entorno Big Data

OBJETIVO

Entender los mecanismos de seguridad que se requieren en una arquitectura Big Data y ubicar soluciones, las cuales permiten desplegarlos, al igual que garantizar la protección de datos alineada a la legislación.

INTRODUCCIÓN

En estos últimos años las empresas tienden a someterse cada vez más a depender de la información a la hora de la toma de decisiones a causa pandemia del COVID-19 se ha incrementado para conseguir que la información sea más accesible y adaptable, haciendo que los instrumentos de integración se han visto acrecentados en los diferentes ámbitos haciendo frente a la avalancha de datos con un nuevo concepto que ha dado en denominarse Big Data.

Por la simple denominación usada se entiende que se trata de grandes volúmenes estructurados y no estructurados de información que no es sencillo tratar con las herramientas y procedimientos tradicionales por su complejidad, velocidad y tamaño de incremento son complicados de procesar mediante herramientas convencionales.

El procesamiento y almacenamiento de información es la clave del Big Data, ya que la gran cantidad de información que se tiene acceso hoy en día no seria relevante de no ser por el desarrollo de sofisticadas herramientas que permitan manejarla sin mencionar la complejidad y variabilidad en el formato de esta información, en donde se almacena en un sitio denominada data lake, en donde el hardware tradicional no está capacitado para el almacenamiento de este volumen descomunal de información, llegando en forma de nube la solución, haciendo estos entornos de almacenamiento a los que se puede acceder desde cualquier sitio con conexión a internet, haciendo que el costo sea mucho más reducido para las corporaciones, dando más espacio de almacenamiento con un costo mucho más accesible, siendo dos factores importantes para el desarrollo del Big Data.

Asignatura	Datos del alumno	Fecha
Seguridad en Bases de Datos y Almacenamiento de Datos Masivos	GRUPO 29	30.07.2022
	Braulio David Velasco Castillo Blanca Paola Toledo Martínez Angel Ramon Paz López	

DEFINICIONES BIG DATA

Se refiere a un término aplicado a conjuntos de datos que superan la capacidad del software habitual para ser capturados, administrado y elaborado en un tiempo razonable. Los tamaños del Big Data se encuentran constantemente en movimiento creciente, actualmente se refieren a conjuntos de datos que van desde 30 – 50 Terabytes a varios Petabytes.

International Business Machines posee la cartera más extensa y profunda del mundo en tecnologías y soluciones Big Data & Analytics, considera que hay Big Data cuando el conjunto de información supera el terabyte de información, es sensible al tiempo, y mezcla información estructurada con no estructurada. Así, su enfoque trata de buscar la forma mejor de aprovechar estos datos, su gestión, su combinación de datos estructurados con los que no lo son, por medio de la aplicación de algoritmos predictivos de comportamiento, y de esta manera permitir la toma de decisiones que añadan valor a las organizaciones.

BIG DATA EN DEFENSA Y SEGURIDAD

En este entorno el propósito es captar y utilizar la gran cantidad de información con el fin de aunar sensores, percepción y decisión en sistemas autónomos, pero la generación de la inmensa cantidad de datos, así como la necesidad de procesarlos y explotarlos de manera eficaz y eficiente es transversal a toda nuestra sociedad, qué duda cabe que el ámbito de la defensa y la seguridad son dos de los entornos donde resulta interesante analizar cómo el Big Data puede aplicarse y ofrecer beneficios.

Para comenzar, se refleja a continuación qué aspectos generales del contexto actual y futuro en defensa y seguridad, presentan un potencial interés desde el punto de vista de Big Data. Tanto el ámbito de la defensa como el de la seguridad están marcados por un enfoque prioritario hacia la prevención. Prevenir siempre es mejor y menos costoso que curar. Sin embargo, la prevención requiere decisiones en ventanas de tiempo muy definidas y con un alto nivel de síntesis de la inmensidad de datos y factores involucrados. Por otro lado, las confrontaciones recientes han visto crecer su grado de complejidad, y con el aumento de la información esto se hará mucho más complejo día a día.

Es termino de complejidad se refiere al mayor número de conectividad de los usuarios y acciones, derivada de la globalización, los escenarios con líneas divisorias muy difusas entre lo civil y lo militar, entornos intensivos en información con creciente mezcla de escenarios virtuales y reales, que sirven para poder trabajar con la creciente complejidad y abundancia de datos, es necesario un mayor enfoque en la comprensión de la situación, especialmente en aquellos ámbitos donde los objetivos son en apariencia de pequeña escala o de carácter indeterminado.

Por lo que no debería ser una sorpresa para el lector que en el ámbito de la seguridad y defensa Big Data se pueda aplicar en áreas de:

- Vigilancia y Seguridad de fronteras
- Sector energético y servicios públicos
- Ciberdefensa / Ciberseguridad
- Lucha contraterrorista y contra crimen organizado
- Industria y manufactura
- Lucha contra el fraude
- Seguridad ciudadana
- Inteligencia militar
- Planeamiento táctico de misiones

Ahora mencionaremos las fases en la que se debe desarrollar la arquitectura Big Data, aunque cada empresa puede tener diferentes procesos para la creación de una arquitectura Big Data, para un mejor desarrollo los proyectos tienen un ciclo de vida y nosotros lo mencionamos en las siguientes fases:

1. **Desarrollar un catálogo de uso:** En todo proyecto Big Data es considerada la fase primordial, el cual permite estudio de casos y los procesos relacionados con el Business Intelligence o el propio Big Data
2. **Planeación de los Datos:** Consiste en la planificación referente a los datos como su extracción, el ciclo de vida y el proceso de análisis, en esta fase se debe planificar también los procesos del mapreduce que es una técnica de procesamiento y un modelo de programación para procesar y analizar grandes cantidades de datos.

3. **Análisis y Diseño:** En esta fase se realizará el debido análisis de la planificación de datos sobre el que y como hacerlo, teniendo claro los objetivos, la forma de vender, los elementos que se necesitarán, los inconvenientes y los puntos positivos de mapreduce y demás herramientas, a la vez se definirán los requisitos para comenzar el diseño de la arquitectura Big Data.
4. **Comprobación del Proceso:** En esta fase se realizan diferentes pruebas para evaluar los procesos y cada equipo involucrado en el proceso como ser ejecutivos, empleados, clientes, técnicos entre otros. Se debe realizar reuniones periódicas y evaluaciones sobre los datos, el trabajo en equipo ya que es un factor importante para el éxito del proyecto.
5. **Integración del Proyecto con los objetivos de la empresa:** El proyecto Big Data que se requiere implementar a una empresa debe ser flexible y coexistir con los objetivos y políticas del negocio, adaptarse a los procesos y a las metas de la empresa.
6. **Ajuste del Impacto:** El proyecto debe adaptarse a las necesidades y a los cambios del mercado, se debe tener una monitorización de los procesos para realizar el análisis debido y los ajustes necesarios.
7. **Buen Marketing:** El trabajo en equipo lo es todo por lo cual todo debe funcionar de la mejor manera vendiendo la imagen del producto de la mejor forma posible sacando provecho de sus utilidades.

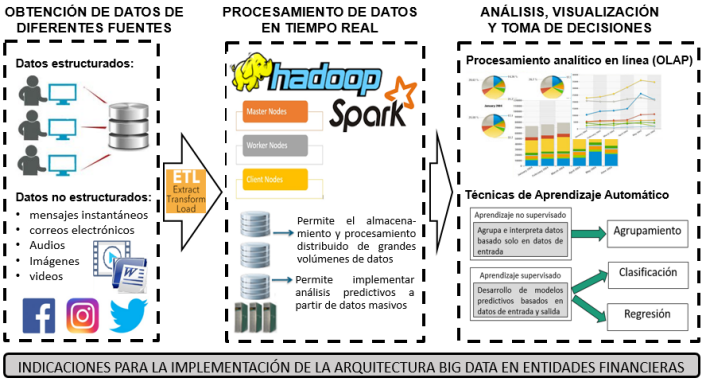
ARQUITECTURA BIG DATA

Esta se enfoca en tratar y analizar grandes volúmenes de información que no pueden ser gestionados de manera convencional done varias organizaciones remedian la problemática de grandes volúmenes de información son recogidos por empresas especializadas, de esta manera los problemas de Big data comparten cinco características, conocidas como las cinco Vs:

- Volumen: Se da a entender como el tamaño de los conjuntos de los datos que normalmente requieren ser almacenados de forma distribuida.
- Variedad: Entra en el contexto que un Big Data se compone de varios tipos diferentes de datos como sonido, texto, imagen y video.
- Veracidad: Se refiere a los sesgos, ruido y anormalidad de los datos.

- Velocidad: Se ocupa del ritmo al que fluyen los datos desde diversas fuentes como dispositivos móviles, Internet de las Cosas (IoT), redes sociales, etc.
- Valor: No solo esta en la posibilidad de recopilar o almacenar información, si no en la capacidad para limpiar y tratar los millones de información que aportamos diariamente.

Así tenemos que Apache Hadoop es la base de todas aquellos framework, en la cual prestaremos interés en un componente de Hadoop denominado Sistema de Archivos Distribuidos de Hadoop (HDFS), la cual describiremos a detalle más adelante. Sin embargo, el framework que se escoge para mitigar el problema presentado es Apache Spark, debido a que se necesita procesos en tiempo real y su velocidad de procesamiento es superior a otras framework.



HADOOP

La arquitectura Hadoop es de código abierto y provee soluciones para manejar Big Data junto con un extenso procesamiento de análisis, permitiendo a los desarrolladores procesar una gran cantidad de datos mientras oculta la complejidad de la ejecución paralela en cientos de servidores en un entorno de la nube.

Este framework está escrito en Java, lo que permite el procesamiento de grandes cantidades de datos a través del clúster usando modelos de programación simple. Así hablando de esta arquitectura, nos centramos primordialmente en el Sistema de Archivos Distribuidos de Hadoop (HDFS) y MapReduce. Estos dos componentes son los más importantes de la arquitectura Hadoop.

Asignatura	Datos del alumno	Fecha
Seguridad en Bases de Datos y Almacenamiento de Datos Masivos	GRUPO 29	30.07.2022
	Braulio David Velasco Castillo Blanca Paola Toledo Martínez Angel Ramon Paz López	

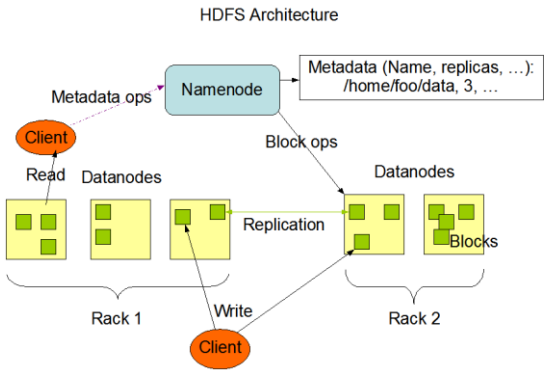
SISTEMA DE ARCHIVOS DISTRIBUIDOS HADOOP (HDS)

El HDFS es un sistema de archivos distribuidos que se puede almacenar datos en varios servidores, de formar paralela el acceso y su respectiva tolerancia a fallos. Una de la diferencia que se tiene este HDFS frente a los otros sistemas de archivos distribuidos, que está diseñado para ser construido a partir de un componente básico de bajo costo que requiere que sea altamente tolerante a fallos.

Este sistema de archivos distribuido está escrito en Java, y se ejecuta sobre el sistema de archivos que usualmente se usa en un equipo computacional y está desarrollado para admitir aplicaciones con grandes cantidades de volúmenes de datos en el orden de terabytes y petabytes.

El funcionamiento que tiene este sistema es dividir un tamaño de bloque típico de 64MB, por lo que es posible almacenar cada fragmento en un nodo diferente otorgando un mayor almacenamiento de archivos, en comparación de un NTFS que es un bloque de 4 KB siendo ineficiente.

La arquitectura que tiene el HDFS es maestro-esclavo. El clúster HDFS consta de un nodo llamado “NameNode” que le pertenece al nodo maestro para administrar el “namespace” del sistema de archivos distribuidos y saber dónde están los bloques de datos almacenados dentro del clúster, además de regular el acceso al archivo por parte de los clientes.



Asignatura	Datos del alumno	Fecha
Seguridad en Bases de Datos y Almacenamiento de Datos Masivos	GRUPO 29	30.07.2022
	Braulio David Velasco Castillo Blanca Paola Toledo Martínez Angel Ramon Paz López	

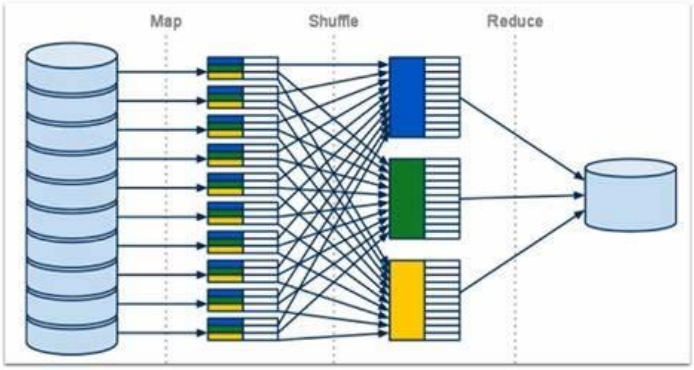
MAPREDUCE

Es una técnica de procesamiento y un modelo de programación para la computación distribuida realizado en Java, que ha sido propuesta por Google como un nuevo modelo de programación para procesar y analizar grandes cantidades de datos.

Su función primordial es dividir y distribuir la carga de trabajo para aumentar la velocidad de cálculo. Esto permite llevar a cabo dos etapas distintas; la operación "Map", que es escrita por el usuario en donde su objetivo es producir una clave o valor intermedio para cada par recibido en la entrada. Luego, la biblioteca MapReduce agrupa todas las claves intermedias asociadas con el mismo valor y las envía a la operación "Reduce".

Como último paso, es fusionar los valores de la misma clave intermedia para crear un valor único asociado con esa clave y devolver una sola clave como salida para cada clave intermedia.

Después del procesamiento mencionado, reúne todos los resultados en un resultado final que será almacenado en el HDFS. Así mismo, MapReduce consta de un servidor maestro al cual se lo denomina JobTracker, que es el responsable de asignar tareas, y una cantidad de servidores esclavos llamados TaskTracker que son los responsables de ejecutar las tareas.



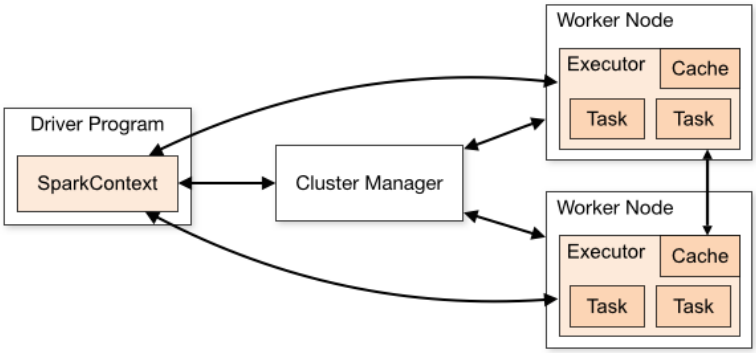
SPARK

Este framework es de código abierto para el procesamiento de Big Data, que permite a los usuarios ejecutar aplicaciones de análisis de datos a gran escala dentro de un clúster.

Siendo este framework uno de los subproyectos de Hadoop desarrollado en 2009 en el AMPLab de UC Berkeley. Así, Spark posee varias ventajas en comparación a otras tecnologías de Big Data del ecosistema de Hadoop como MapReduce. Primero, ofrece un completo y unificado framework para direccionar a las necesidades del tratamiento de datos.

Spark está desarrollado en Scala que es programación funcional adecuada para más distribuidos. Gracias a ello, Spark permite que las aplicaciones en los clústeres de Hadoop funcionen unas 100 veces más rápido en la memoria, y unas 10 veces más rápido en el disco de almacenamiento.

Esto llega a ser posible debido a la reducción del número de operaciones de lectura y escriturad el disco donde se almacena los datos intermedios de procesamiento en la memoria del clúster, como se puede observar en la siguiente imagen donde se expresa el tiempo de ejecución al aplicar una regresión logística tanto en Hadoop como en Spark.

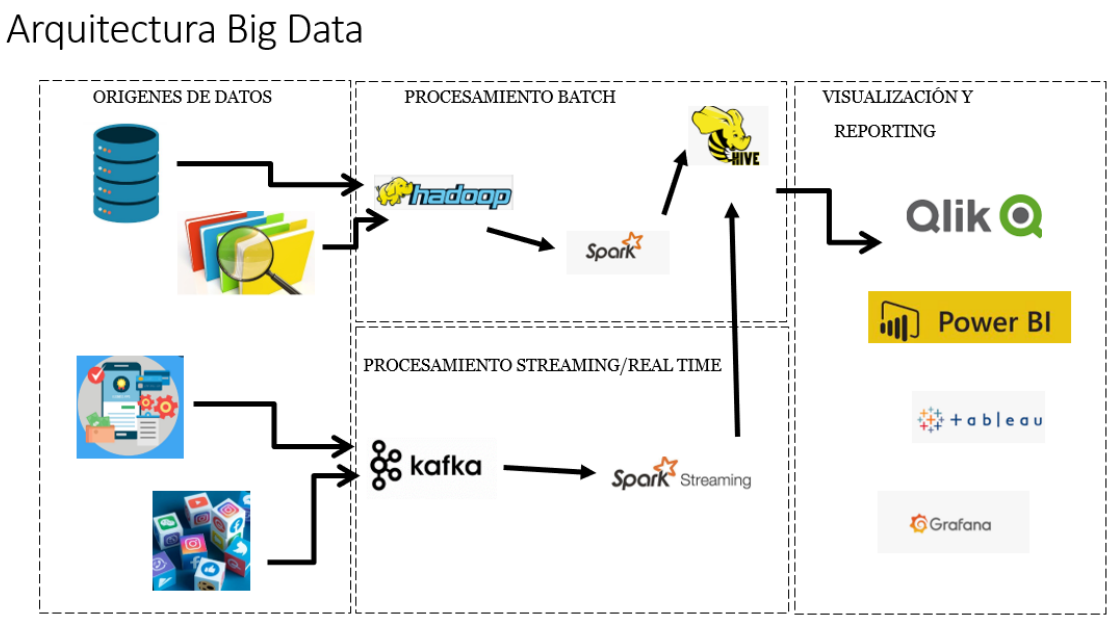


A parte de las operaciones de “Map” y “Reduce”, Spark incorpora consultas SQL, funciones de transmisión de datos, Machine Learning y algoritmos de gráfico. Los desarrolladores pueden usar estas funciones en forma independiente o dentro de una cadena de procesamiento complejo.

Como podemos observar en la siguiente imagen, está representando los componentes que contiene Spark, los cuales mencionaremos a continuación:

- Spark SQL: Puede mostrar conjuntos de datos de Spark por medio de la JDBC API y ejecutar las consultas convenientes del tipo SQL utilizando herramientas empresariales inteligentes y con su visualización tradicional.
- Spark Streaming:Se puede usar para el procesamiento de flujo de datos en tiempo real.
- Spark MLlib:MLlib es un framework de Machine Learning distribuido en Spark, que proporciona múltiples tipos de algoritmos de aprendizaje automatizados que incluyen clasificaciones, regresiones, clustering y filtrado colaborativo, y también proporciona funcionalidades como evaluación de modelos e importación de datos.
- GraphX:Es una nueva API para el tratamiento de gráficos que incluye la ejecución en paralelo.

En el presente diagrama se tiene una arquitectura de Big Data en entorno productivo y en donde podemos observar como se realiza el procesamiento de los datos desde el origen de los datos hasta la visualización y reporteo para consulta del negocio.



Asignatura	Datos del alumno	Fecha
Seguridad en Bases de Datos y Almacenamiento de Datos Masivos	GRUPO 29	30.07.2022
	Braulio David Velasco Castillo Blanca Paola Toledo Martínez Angel Ramon Paz López	

SOLUCIONES DE SEGURIDAD PARA BIG DATA

APACHE SENTRY: La herramienta permite autorización granular y un control de acceso basado en roles obteniendo una autorización unificada, Se adapta a herramientas como Apache Hive, HDFS, KAFKA también puede integrarse con otras herramientas del ecosistema Hadoop.

CLOUDERA SEARCH: La herramienta permite tener una solución integral en toda la arquitectura en temas de auditoria de datos y procesos adecuándose al cumplimento de las normas para Hadoop, brinda auditoria y permite obtener un historial, ayuda a los usuarios a identificar el origen, uso y el impacto de los datos facilitando el gobierno de los datos.

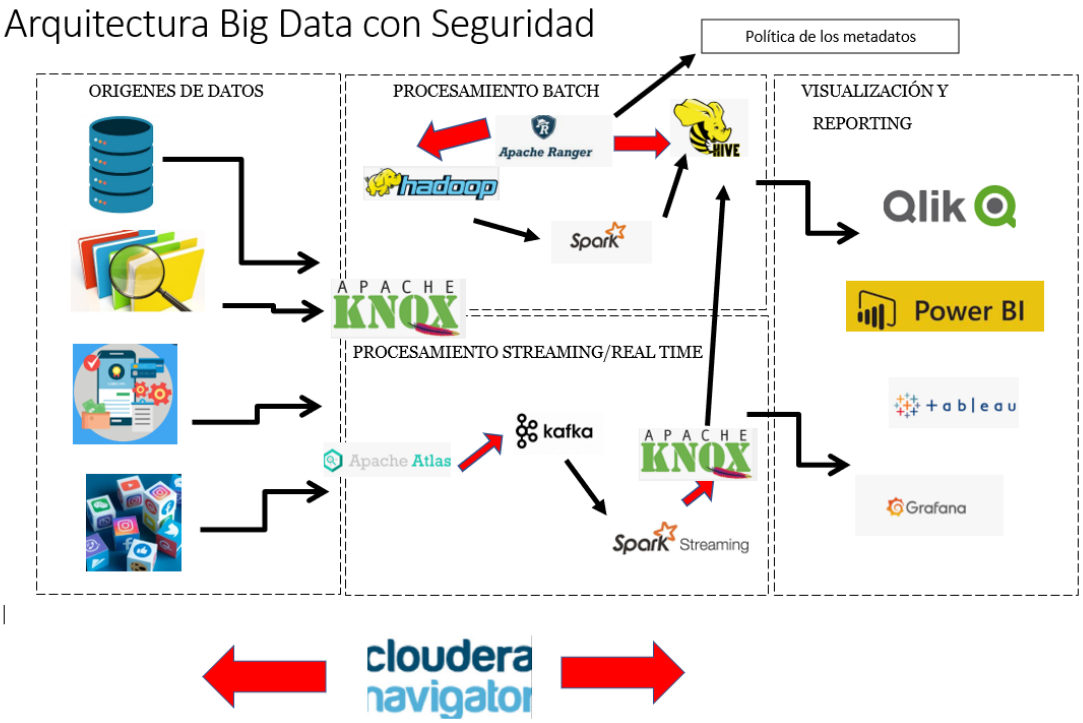
APACHE RANGER: La herramienta como lo indica su página brinda las siguientes facilidades:

- Administración de seguridad centralizada para administrar todas las tareas.
- Autorización detallada para realizar una acción y/u operación específica con el componente/herramienta hadoop.
- Estandarice el método de autorización en todos los componentes de Hadoop.
- Soporte mejorado para diferentes métodos de autorización: control de acceso basado en roles, control de acceso basado en atributos, etc.
- Centralizar la auditoría del acceso de los usuarios y las acciones administrativas (relacionadas con la seguridad) dentro de todos los componentes de Hadoop.

APACHE KNOX: La herramienta es una aplicación Gateway que permite la interacción con las API REST y las implementaciones de Apache Hadoop. Es un proxy que funciona de manera inversa contando con el cumplimiento de las políticas a través de los proveedores y de los servicios. El Knox Gateway permite la protección de los clústeres y que el consumidor de la API tenga un único punto de conexión para el acceso a los servicios necesarios.

APACHE ATLAS: La herramienta permite un conjunto de servicios de gobernanza para que las organizaciones cumplan los requisitos de cumplimiento de seguridad de Hadoop permitiendo integrar el ecosistema con toda la organización. Permite:

- Tipos predefinidos para varios metadatos de Hadoop y no Hadoop
- Capacidad para definir nuevos tipos de metadatos a gestionar
- Las API de REST para trabajar con tipos e instancias permiten una integración más fácil



Asignatura	Datos del alumno	Fecha
Seguridad en Bases de Datos y Almacenamiento de Datos Masivos	GRUPO 29	30.07.2022
	Braulio David Velasco Castillo Blanca Paola Toledo Martínez Angel Ramon Paz López	

CONCLUSIONES

Actualmente, en un mundo cada vez más digitalizado, diariamente se crean grandes volúmenes de datos en diferentes entornos tantos laborables, educativos, científicos, etc., por lo que es necesario crear arquitecturas que procesen, almacenen y visualicen enormes cantidades de datos. Estas arquitecturas deben responder a las cuatro V's (Volumen, Variedad, Veracidad, Velocidad). Durante el desarrollo del trabajo se pueden observar varias tecnologías Big Data específicas para cada uso, con sus ventajas e inconvenientes bien definidas.

Hasta aquí se ha intentado dar una visión de lo que se puede entender por Big Data, qué funciones abarca y cuáles son las herramientas y recursos disponibles. Acorde con la gran cantidad de información que hoy en día se maneja en Internet, la imagen es el principal medio de comunicación, ya sea fija o en movimiento, todo lo envuelve, y así se ha incidido en las técnicas más básicas del tratamiento de imágenes y la vertiente matemática necesaria. Ciñéndonos a los entornos de defensa y seguridad se han incluido un conjunto de aplicaciones posibles, así como un análisis de cómo pueden evolucionar estas aplicaciones.

El utilizar la tecnología Apache Sparkpor múltiples ventajas como el paralelismo simplificado que se utiliza con un módulo llamado Spark Streaming, en el que crea varias particiones del "Resilient Distributed Dataset" (RDD) trabajando cada una en un nodo esclavo o worker. Otro punto de vista es la alta escalabilidad que se presenta, así como el procesamiento complejo de los datos con funciones de alto nivel. Esto nos conlleva a un alto rendimiento a comparación de otras tecnologías de datos distribuidos, y un procesamiento de datos más rápido que Apache Hadoop MapReduce. Además, es respaldado en la integración de esta tecnología en grandes compañías como IBM, Amazon, Microsoft y Google.

Asignatura	Datos del alumno	Fecha
Seguridad en Bases de Datos y Almacenamiento de Datos Masivos	GRUPO 29	30.07.2022
	Braulio David Velasco Castillo Blanca Paola Toledo Martínez Angel Ramon Paz López	

REFERENCIAS

Análisis y propuesta de arquitectura para garantizar seguridad en entornos Big Data. (s. f.). Recuperado 28 de julio de 2022, de https://repositorio.uam.es/handle/10486/681157

Asignatura	Datos del alumno	Fecha
Seguridad en Bases de Datos y Almacenamiento de Datos Masivos	GRUPO 29	30.07.2022
	Braulio David Velasco Castillo Blanca Paola Toledo Martínez Angel Ramon Paz López	

HOJA DE CONTROL DE ACTIVIDAD GRUPAL

Cada integrante debe llenar e incluir al final de su documento el siguiente registro.

Hoja de control de actividad grupal **<Nombre y apellidos del primer miembro del equipo >**			
	Marcar con una **X** lo que proceda		
Asistencia a reuniones de equipo	Asistencia a una sesión o ninguna	Asistencia a dos sesiones	Asistencia a tres sesiones
- Integrante 1			x
- Integrante 2			x
- Integrante 3			x
Tareas o entregas a realizadas	Ninguna o una tarea	Dos tareas	Tres tareas
- Integrante 1			x
- Integrante 2			x
- Integrante 3			x

Asignatura	Datos del alumno	Fecha
Seguridad en Bases de Datos y Almacenamiento de Datos Masivos	GRUPO 29	30.07.2022
	Braulio David Velasco Castillo Blanca Paola Toledo Martínez Angel Ramon Paz López	