UNIVERSIDAD TÉCNICA DEL NORTE

INSTITUTO DE POSTGRADO

MAESTRÍA EN TELECOMUNICACIONES

Proyecto del Trabajo de Titulación previo a la obtención del Título de Magíster en Telecomunicaciones

TEMA:

“EVALUACIÓN DE UNA ARQUITECTURA DE BIG DATA PARA LA RED MÓVIL 5G A NIVEL DE LA CAPA INGESTIÓN UTILIZANDO APLICACIONES DE RECOLECCIÓN DE DATOS”

AUTORA:

SAMANTHA MARISOL MESA TAPIA

DIRECTORA:

MSC. SILVIA DIANA MARTÍNEZ MOSQUERA

IBARRA - ECUADOR

2021

APROBACIÓN DEL TUTOR

Yo, **Silvia Diana Martínez Mosquera**, certifico que la estudiante **Samantha Marisol Mesa Tapia** con Cédula **N°100339735-1** ha elaborado bajo mi tutoría la sustentación del trabajo de grado titulado: **"Evaluación de una arquitectura de Big Data para la red móvil 5G a nivel de la capa ingestión utilizando aplicaciones de recolección de datos**".

Este trabajo se sujeta a las normas y metodologías dispuestas en el reglamento del título a obtener, por lo tanto, autorizo la presentación a la sustentación para la calificación respectiva.

Ibarra, 07 de octubre de 2021

MSc. Silvia Diana Martínez Mosquera

Tutora

C.I.: 1718478603

UNIVERSIDAD TÉCNICA DEL NORTE
INSTITUTO DE POSTGRADO
BIBLIOTECA UNIVERSITARIA

AUTORIZACIÓN DE USO Y PUBLICACIÓN A FAVOR DE LA UNIVERSIDAD TÉCNIA DEL NORTE

1. IDENTIFICACIÓN DE LA OBRA

En el cumplimiento del Art. 144 de la Ley de Educación Superior, hago la entrega del presente trabajo a la Universidad Técnica del Norte para que sea publicado en el Repositorio Digital Institucional, para lo cual pongo a disposición de la siguiente información:

DATOS DE CONTACTO	
CÉDULA DE IDENTIDAD:	1003397351
APELLIDOS Y NOMBRES:	Mesa Tapia Samantha Marisol
DIRECCIÓN:	San José de Cananvalle
EMAIL:	smmesat@utn.edu.ec
TELÉFONO MÓVIL:	0999063232
DATOS DE LA OBRA	
TÍTULO:	"Evaluación de una arquitectura de Big Data para la red móvil 5G a nivel de la capa ingestión utilizando aplicaciones de recolección de datos"
AUTOR:	Mesa Tapia Samantha Marisol
FECHA:	22 de octubre de 2021
SOLO PARA TRABAJOS DE GRADO	
PROGRAMA:	( ) PREGRADO (X) POSGRADO
TÍTULO POR EL QUE OPTA:	Magister en Telecomunicaciones
ASESOR/DIRECTOR:	MSc. Silvia Diana Martínez Mosquera

2. CONSTANCIAS

La autora Samantha Marisol Mesa Tapia, manifiesta que la obra de objeto de la presente autorización es original y se la desarrolló, sin violar derechos de autor de terceros, por lo tanto, la obra es original y que es el titular de los derechos patrimoniales, por lo que asume la responsabilidad sobre el contenido de esta y saldrá en defensa de la Universidad en caso de reclamación por parte de terceros.

Ibarra a los 22 días del mes de octubre del año 2021

Samantha Marisol Mesa Tapia

C.I.: 1003397351

DEDICATORIA

A mis padres Patricio y Marisol, una bendición de Dios en este mundo, dos grandes seres humanos, luchadores, trabajadores e incansables, que me han enseñado el valor de la vida, de la familia, de lo sencillo, el amor al prójimo y valores que solo pueden ser fundados en el hogar.

A mi hermano Israel, porque hemos crecido juntos y a mi hermano Ariel, nuestro ángel que está en el cielo, son la muestra del amor inquebrantable que existe entre hermanos.

AGRADECIMIENTOS

A Dios, por nunca soltar mi mano, ser la base de mi vida y por abrir nuevas puertas para mi crecimiento profesional.

A mis padres, por su presencia y su apoyo en este trayecto, son el motor para que este logro sea posible.

A mi tutora, Dianita Martínez, por ser la principal guía para el desarrollo de este proyecto de titulación, por su apoyo constante con su conocimiento, su paciencia y su amistad.

ÍNDICE DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

UNIVERSIDAD TÉCNICA DEL NORTE

INSTITUTO DE POSGRADO

PROGRAMA DE MAESTRÍA

"EVALUACIÓN DE UNA ARQUITECTURA DE BIG DATA PARA LA RED MÓVIL 5G A NIVEL DE LA CAPA INGESTIÓN UTILIZANDO APLICACIONES DE RECOLECCIÓN DE DATOS"

Autora: Samantha Marisol Mesa Tapia
Tutora: Msc. Silvia Diana Martínez Mosquera
Año: 2021

RESUMEN

La evolución de las redes móviles ha traído consigo cambios relevantes que han permitido mejorar su servicio en gran manera, cada una ha manejado su estándar y su tecnología, así también la velocidad en la que trabajan ha ido incrementando al igual que el número de usuarios y por ende la transmisión de datos. Esto a su vez plantea nuevos desafíos, ya que la red móvil 5G pretende llegar a una velocidad mucho mayor que las ya conocidas con una transmisión de un gran volumen de datos que en las tecnologías tradicionales no se pueden manejar correctamente, se requiere una tecnología más robusta como Big Data para recolectar esa gran cantidad de datos de la red móvil 5G para luego poder procesarlos y analizarlos, en ese sentido, la identificación de aplicaciones adecuadas en un contexto de 5G y Big Data es un problema abierto. Es por ello que se realiza la evaluación del desempeño de una arquitectura de Big Data para 5G a nivel de la capa de ingestión, usando tres herramientas de recolección de datos que trabajan con el formato de archivos que se transmiten en la red móvil 5G y que permiten analizar el tiempo de procesamiento en base al tamaño de los datos generados; para la selección de las herramientas de ingestión de datos se realizó un análisis en base a algunos indicadores tanto de funcionalidad como de rendimiento, esto permitió tener una visión más clara sobre las herramientas a utilizarse, y con las cuales se realizó el proceso de pruebas utilizando varias muestras con diferentes tamaños y poder estresar al sistema para evaluar el desempeño de cada una de ellas en la capa de ingestión de datos y posterior a eso recomendar la herramienta más eficiente para trabajar con archivos generados en una red móvil 5G.

Palabras clave: 5G, Big Data, ingestión.

UNIVERSIDAD TÉCNICA DEL NORTE

INSTITUTO DE POSGRADO

PROGRAMA DE MAESTRÍA

"EVALUACIÓN DE UNA ARQUITECTURA DE BIG DATA PARA LA RED MÓVIL 5G A NIVEL DE LA CAPA INGESTIÓN UTILIZANDO APLICACIONES DE RECOLECCIÓN DE DATOS"

Autora: Samantha Marisol Mesa Tapia
Tutora: Msc. Silvia Diana Martínez Mosquera
Año: 2021

ABSTRACT

The evolution of mobile networks has brought with it relevant changes that have made it possible to improve their service in a great way, each one has managed its standard and its technology, as well as the speed at which they work has been increasing as well as the number of users and hence the transmission of data. This in turn poses new challenges, since the 5G mobile network aims at a much higher speed than known with a transmission of a large volume of data that in traditional technologies cannot be handled correctly, a more robust technology is required such as Big Data to collect that large amount of data from the 5G mobile network and then be able to process and analyze it, in that sense, the identification of suitable applications in a 5G and Big Data context is an open problem. That is why the performance evaluation of a Big Data architecture for 5G is carried out at the level of the ingestion layer, using three data collection tools that work with the format of files that are transmitted in the 5G mobile network and that allow to analyze the processing time based on the size of the data generated; For the selection of the data ingestion tools, an analysis was carried out based on some indicators of both function and performance, this has a clearer vision about the tools to be used, and with which the testing process was carried out using various samples with different sizes and to be able to stress the system to evaluate the performance of each one of them in the data ingestion layer and after that recommend the most efficient tool to work with files generated in a 5G mobile network.

Keywords: 5G, Big Data, ingestion.

CAPÍTULO I

El Problema

Desde inicios de la existencia de las comunicaciones en el mundo, han venido cambiando de acuerdo a las necesidades de las generaciones de la población, conllevando a nuevas formas de evolución de las telecomunicaciones que generan cambios en los actuales paradigmas, teniendo la expectativa de tener mejores condiciones en la comunicación, se han dado cambios tanto en la naturaleza del sistema, la velocidad, la tecnología y la frecuencia ya que cada generación tiene algunos estándares, capacidades técnicas y nuevas características que la diferencian de la anterior. Por consiguiente, se desarrollan paralelamente innovaciones tanto del hardware como del software de los componentes que intervienen en este tipo de tecnologías (González & Salamanca, 2016).

La primera generación (1G) utilizaba tecnología analógica, en la segunda generación (2G) los teléfonos utilizaban tecnología digital y fue con esta red que el uso de estos dispositivos pasó a ser accesible para todos, se contaba con la comunicación por voz, pero no era la adecuada para trabajar con datos pues la velocidad manejada entre la red y el dispositivo era, en primera instancia, de 900 bits por segundo, en la tercera generación (3G) se permiten las videollamadas y la conexión de datos se da a 384000 bits por segundo, en la cuarta generación (4G) se tiene una red mejorada con mayor cobertura, su gran diferencia con las generaciones anteriores es la capacidad para proveer velocidades de acceso mayores de 100 Mbps en movimiento y 1 Gbps en reposo, lo que permite ofrecer servicios de cualquier clase en cualquier momento y en cualquier lugar (Díaz, 2016).

Pese a estos importantes avances, tales como la transmisión de datos con mayor velocidad y el uso de diferentes aplicaciones en un solo dispositivo se puede ver que cada vez el número de usuarios se incrementa en gran manera, tomando en cuenta también que un mismo usuario puede hacer uso de varios dispositivos lo que hace que en la red móvil se genere una gran cantidad de datos, que en algunos casos estos datos no tienen una estructura definida como son los datos generados por los usuarios en redes sociales, video streaming, notas de voz o imágenes.

Otro escenario a considerar es la gran cantidad de datos que generan los sensores que utilizan Internet de las Cosas de sus siglas en ingles IOT[1], lo que hace necesario que existan tratamientos adecuados para toda esta generación masiva de datos, mismo que las tecnologías tradicionales no cuentan y tienen que cubrir este requerimiento.

Al hablar de la quinta generación (5G), se pretende que alcance incluso los 10 Gbps con una latencia de 1ms, misma que permita la transmisión de un gran volumen de datos, hasta 100 dispositivos más conectados por unidad de área (en comparación con las redes 4G LTE[2]), con reducción del consumo de energía del 90%, lo que hace necesario contar con una tecnología que sea un aporte a la red móvil 5G y que permita dar tratamiento a toda esa gran cantidad de información como es el caso de Big Data, que puede trabajar con diferentes tipos de datos como estructurados y no estructurados, además de un manejo a altas velocidades y análisis para darles valor, identificar patrones y generar aportes para las empresas que lo implementen, lo que conlleva a la toma de decisiones importantes (MENDOZA, 2016).

De acuerdo con algunos estudios previos realizados en (Kamakhya Narain Singh, 2018), las arquitecturas de Big Data se componen de varias capas, en donde se ve la importancia de la capa ingestión o también llamada capa de introducción de datos. Las soluciones actuales no usan paradigmas de Big Data en la capa Ingestión, usan herramientas tradicionales, las cuales con 5G deben ser más robustas para poder evaluar su desempeño en términos de eficacia. En este sentido, la identificación de las aplicaciones adecuadas en un contexto de 5G y Big Data es una clara necesidad y se considera un problema abierto.

Objetivos de la investigación

Objetivo general

Evaluar el desempeño de una arquitectura de Big Data para 5G a nivel de la capa de ingestión, usando diferentes aplicaciones de recolección de datos.

[1] IOT: Internet of Things
[2] LTE: Long Term Evolution

Objetivos específicos

- Analizar el formato de los datos de una red móvil 5G a ser procesados en la capa de Ingestión para determinar las aplicaciones de recolección de datos que trabajen con el mismo.
- Evaluar tres aplicaciones de recolección de datos que permitan analizar el tiempo de procesamiento en base al tamaño de los datos generados.
- Realizar pruebas de estrés del sistema para evaluar el desempeño de la capa de ingestión con cada herramienta seleccionada.
- Identificar la aplicación más eficiente en base de los resultados obtenidos en el proceso de recolección de datos 5G en la capa ingestión.

Justificación

En esta nueva tecnología de comunicaciones las características principales de crecimiento de usuarios y dispositivos de forma masiva en la red móvil 5G, crean la necesidad de contar con una tecnología que tenga la capacidad de procesamiento, velocidad, y a la vez pueda tratar con distintos tipos de datos, ya que se logrará de forma autónoma realizar desde lo más sencillo como ir a un lugar a otro hasta operaciones médicas complejas a distancia, se pueden implementar ciudades inteligentes, realidad virtual, vehículos terrestres y aéreos automáticos, etc., se haría realidad mucho de lo que ahora se conoce como ciencia ficción (Barreno et al., 2016)

En efecto, al trabajar en conjunto Big Data y 5G, se genera una gran expectativa, ya que se generan grandes innovaciones que abarquen lo real y lo virtual con velocidades dentro de los zettabytes. Por tanto, el almacenamiento e ingestión de los datos representa una necesidad y un gran desafío para las tecnologías existentes.

El Plan Nacional de Desarrollo 2017-2021 plantea en su objetivo 5 Impulsar la productividad y competitividad para el crecimiento económico sostenible de manera redistributiva y solidaria a través de una base de recurso naturales renovables y no renovables, incrementar las exportaciones agropecuarias y agroindustriales en al menos el 33% a 2021; incrementar de 4,6 al 5,6 el índice de

Desarrollo de Tecnologías de Información y Comunicación al 2021, En consecuencia, esta nueva arquitectura productiva favorecerá: El uso de tecnologías aplicadas al incremento de la productividad.

En este proyecto de investigación se plantea realizar la **EVALUACIÓN DE UNA ARQUITECTURA DE BIG DATA PARA LA RED MÓVIL 5G A NIVEL DE LA CAPA INGESTIÓN UTILIZANDO APLICACIONES DE RECOLECCIÓN DE DATOS.** Además, se busca aportar al cumplimiento de la misión de la Universidad Técnica del Norte que genera, fomenta, y ejecuta procesos de Investigación, de transferencia de saberes, de conocimientos científicos, tecnológicos y de innovación.

Finalmente, este trabajo contribuye a la línea de Investigación "Innovación tecnológica y productos de telecomunicación" que está alineada a las líneas institucionales y al programa de postgrado de la Universidad Técnica del Norte

CAPITULO II: MARCO REFERENCIAL

Antecedentes

Para el desarrollo del presente trabajo se ha tenido en cuenta la importancia del desarrollo previo de esta tecnología 5G y como fue constituida para actualmente dar paso a grandes avances en comunicación móvil e integración de servicios IT[3] pese a que aún no se encuentre estandarizada. Se prevé que para el año 2020 ya se encuentre en total funcionamiento y a partir de esto se pueda convertir en la tecnología habitual de todos los usuarios en general. (Vera Cárdenas, 2018)

Posterior a esto en países como Suecia, España, Finlandia e incluso China han apostado por inversiones en el estudio y desarrollo de esta nueva generación de telefonía móvil. De forma particular en Europa en el año 2013 se destinaron alrededor de €50 millones para la investigación de 5G y que este pueda ser implementado en el año 2020, esto bajo la tutela de la Unión Europea.(Vera Cárdenas, 2018)

[3] IT: Information Technology

En Suecia la multinacional Ericsson ha realizado constantes avances en su infraestructura para la implementación de 5G, dando demostraciones de la funcionalidad en términos de velocidad y latencia en conjunto con operadoras en diferentes países, especialmente, a nivel de Latinoamérica. En el año 2015 Ericsson realizó mejoras en el software de sus productos de sistema de radio para que estos puedan soportar la nueva tecnología 5G. (Reichert, 2018)

Por otra parte, en el MWC[4] 2018, encuentro institucional que se lleva a cabo cada año para discutir y presentar los avances tecnologías en materia de comunicación móvil y desarrollo de esta tuvo como principal tópico la tecnología 5G y como esta abrirá las puertas a nuevas etapas de conectividad que hasta el momento se habían considerado. Dentro de este congreso importantes empresas como Ericsson e Intel recalcaron que 5G no es solo una idea más, sino más bien, es ya una necesidad para poder manejar y canalizar el creciente flujo de datos en los dispositivos y que estos puedan contar con un servicio adecuado. La nueva tecnología en desarrollo va más allá de los autos y casas inteligentes, con 5G la integración se podrá realizar a muchas más aplicaciones y dispositivos al mismo tiempo, puesto que con una mayor capacidad en velocidad y menor latencia el servicio será más eficiente y por ende consumirá menos energía, lo que garantiza estabilidad y escalabilidad en la conectividad. Se tiene en mente que para la próxima edición del MWC ya se tengan definidos estándares para el uso de 5G en términos de infraestructura y uso de espectro, así como también, se llegue a un consenso del ancho de banda asignado para que se puedan alcanzar las velocidades esperadas para la nueva tecnología (MWC, 2018).

Por otro lado, también es importante mencionar la necesidad de las tecnologías que posteriormente se han definido como Big Data, un ejemplo claro es la evolución de Google, a medida que iba creciendo necesitaba más almacenamiento y procesar grandes volúmenes de datos, otra de las grandes empresas que impulsa el desarrollo de tecnologías Big Data como es Amazon también cuenta con su sistema de almacenamiento de datos masivos (Niño & Illarramendi, 2015).

En paralelo a estos avances, cada vez gana más protagonismo el impulso del Big Data desde las compañías tecnológicas que surgen con la eclosión de la web y las redes sociales, como Facebook, Twitter o Instagram, considerando que los datos generados en estas aplicaciones se transmiten también por la red móvil. Estas empresas parten de una necesidad de negocio similar a la que motivó

[4] MWC: Mobile World Congress

a Google a iniciar el desarrollo de tecnologías Big Data, que necesitan de herramientas específicas para sus aplicaciones particulares, donde se identifiquen los principios y elementos fundamentales con los que debe contar un sistema de este tipo, y cómo se interrelacionan para responder a diferentes necesidades de procesamiento de datos, como por ejemplo el manejo de datos masivos almacenados como grafos, para procesar las conexiones de usuarios en una red social, o el procesamiento de flujos de datos masivos al mismo tiempo que se van generando en la red, da lugar a ofrecer un análisis en tiempo real y a la vez asegurar objetivos como la escalabilidad o la tolerancia a fallos.(Niño & Illarramendi, 2015)

Marco teórico

Tecnología 5G

Si se hace un breve análisis de la evolución de las redes móviles sin duda se evidencian los grandes avances que hasta la actualidad se han tenido, pero es importante mencionar que la tecnología que más expectativas está levantando es la llegada del 5G ya que esta tecnología promete hacer posible tener un mundo hiperconectado en donde la mayoría de los objetos de uso cotidiano estarán conectados entre sí también con las personas. De acuerdo a lo que se menciona en (GSMA, 2020) el mercado móvil de la región alcanzará varios hitos importantes en los próximos cinco años, entre ellos 15 millones de conexiones móviles 5G en 2022, esto representa un reto grande para las operadoras que existen en el mercado, para lograr eso la tecnología 5G contará con mayor ancho de banda, menor latencia y mayor capacidad para conectar muchos dispositivos, algo que se debe destacar y se debe tomar en cuenta por su importancia es la eficiencia energética que sin duda es indispensable para que pueda ser posible el desarrollo tecnológico. (Gutiérrez Álvaro, 2019)

Si se analiza este contexto y la experiencia de las generaciones anteriores de las redes móviles, es una responsabilidad grande que el 5G sea un éxito y se desarrolle de una manera sostenible, así como lo fue el 2G, actualmente lo está siendo el 4G y así evitar los errores que se dieron en los comienzos inciertos del 3G, se deben tomar en cuenta varios factores, entre ellos qué elementos deben considerarse para funcione el 5G y se ponga en marcha (Gutiérrez Álvaro, 2019).

Aparte del notable incremento en la velocidad de conexión y de la transmisión de datos a través del servicio, también está la optimización del espectro electromagnético, ya que se convierte en casi fundamental, el uso de antenas MIMO[5], estas permiten a los dispositivos que trabajen en múltiples frecuencias de forma simultánea es decir que tienen múltiples entradas y múltiples salidas que permiten optimizar la comunicación y por ende las velocidades de transmisión y recepción de la información (Jaramillo et al., 2017).

Arquitectura 5G

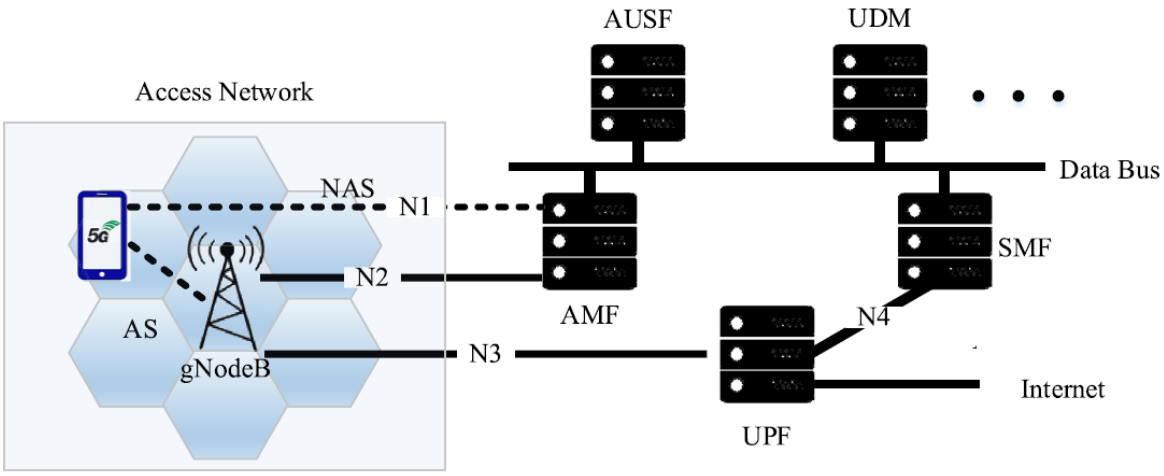
La arquitectura del sistema 5G se define para admitir la conectividad de datos y los servicios que permiten que las implementaciones utilicen técnicas como, por ejemplo, Virtualización de funciones de red y redes definidas por software. (3rd Generation Partnership Project, 2021). La tecnología 5G emplea una arquitectura más inteligente, con redes de acceso por radio RAN[6]. La tecnología 5G lidera el camino hacia una red RAN virtual, flexible y descompuesta con interfaces nuevas que crean puntos de acceso de datos adicionales, y esto se basa en dar solución a la interferencia de la señal en las redes inalámbricas porque afecta negativamente a la cobertura de transmisión y capacidad, lo que limita el rendimiento general de la red. (SOLUTIONS, 2021)

la red 5G presenta algunas características nuevas a las anteriores, como control y separación del plano de usuario (CUPS), tiene una arquitectura basada en servicios (SBA). Las funciones de red en 5GC se encuentran conectados a través de un bus de datos.

A continuación, en la figura 1 se tiene una explicación que permite comprender la arquitectura de 5G, está presenta de forma simplificada y se compone de: equipo de usuario UE[7], red de acceso 5G 5G-AN[8], red central 5G y la red de datos. Como se observa en la figura, la red 5G presenta algunas características nuevas, como control y separación del plano de usuario CUPS[9], arquitectura basada en servicios (SBA por sus siglas en inglés). Las funciones de la red central en 5G están conectados a través de un bus de datos. (Hu et al., 2019)

[5] MIMO: Multiple Input - Multiple Output
[6] RAN: Radio access network
[7] UE: User equipment
[8] 5GAN: Access Network 5G
[9] CUPS: Control User Plane Separation

Figura 1 *Arquitectura de la red 5G*



Fuente: (Hu et al., 2019)

Equipo de Usuario

El UE generalmente consta de dos partes, es importante conocer qué es cada uno de estos componentes el equipo móvil (ME por sus siglas en inglés) y la tarjeta o módulo de identidad de suscriptor universal (USIM por sus siglas en inglés) , el equipo móvil es un dispositivo de hardware que permite la comunicación del usuario, que incluye un CPU, un chip de banda base, una pantalla, una batería, etc., el equipo móvil almacena información específica sobre la Identidad de Equipo Móvil (IMEI por sus siglas en inglés), que determina de forma exclusiva la identidad del dispositivo. La tarjeta o módulo de identidad de suscriptor universal es un módulo de identidad de usuario universal emitido por operadores y es la que almacena información como el identificador permanente de suscripción del usuario, la clave raíz y la clave pública del operador. Esta información puede usarse para identificar de forma exclusiva a un suscriptor legítimo y completar la autenticación mutua cuando el equipo de usuario requiera el acceso a la red 5G. (Hu et al., 2019)

5G-AN

La red 5G incluye múltiples métodos de acceso, como el acceso 3GPP y acceso no 3GPP. Es similar a red celular heredada, pero con una gran cantidad de gNodeB[10] que son las estaciones base de 5G y sustituyen a las eNodeB[11] de LTE, la forma de la red de acceso 5G, se divide el área geográfica en celdas hexagonales, como resultado de qué el equipo de usuario puede acceder a la red 5G en cualquier momento y en cualquier lugar. Cuando el equipo de usuario intenta acceder al sistema 5G, el gNodeB tiene para asignarle recursos de radio. Después de la conexión en donde se realiza el control de recursos de radio (RRC por sus siglas en inglés) el equipo de usuario puede continuar estableciendo la conexión de estrato sin acceso (NAS por sus siglas en inglés) que es un conjunto de protocolos que utilizan para transmitir señales que no son de radio entre el equipo de usuario y la entidad de gestión de movilidad (Hu et al., 2019).

AMF

Esta es la función de gestión de acceso y movilidad (AMF por sus siglas en inglés), es la terminación de señalización NAS[12] en el lado de la red, que es responsable de la gestión de registro, gestión de conexión, gestión de accesibilidad, gestión de movilidad en el sistema 5G, así como cifrado de mensajes NAS, protección de integridad. En comparación con 4G LTE, 3GPP separa el acceso de la función de control y gestión de movilidad de (MME[13] por sus siglas en inglés) al formar el AMF[14] en la red 5G. Al mismo tiempo, la sesión la función de gestión se separa de la función de control y gestión de movilidad para formar la función de gestión de sesión (SMF[15] por sus siglas en inglés) en la red 5G (Hu et al., 2019).

AUSF

Esta es la función del servidor de autenticación (AUSF[16] por sus siglas en inglés) es la principal responsable para autenticación del acceso 3GPP y no-3GPP o acceso no confiable (Hu et al., 2019).

[10] gNodeB: Estaciones base de la red móvil 5G
[11] eNodeB: Estaciones base de la red móvil 4G
[12] NAS: Non-Access Stratum
[13] MME: Mobility Management Entity
[14] AMF: Access and Mobility Management Function
[15] SMF: Session Management Function
[16] AUSF: Authentication Server Function

UDM

La función de gestión de datos unificada (UDM[17] por sus siglas en inglés) es la principal responsable de la generación de credenciales de autenticación, el almacenamiento y la gestión de toda la información de identidad del usuario (SUPI por sus siglas en inglés) en el sistema 5G, así como descifrado del identificador permanente de suscripción único (Hu et al., 2019).

Para monitorear y medir continuamente el desempeño de la red se tienen los datos de gestión del rendimiento (PM[18] por sus siglas en inglés), estos datos se recopilan de cada elemento de la red 5G en un sistema de gestión de red (NMS por sus siglas en inglés),los componentes de mediación son los responsables de los datos PM por lo que estos datos irían desde el equipo de usuario, la red de acceso, la red central y la red de datos hacía el sistema de gestión de red (Martinez et al., 2020).

Casos de uso de 5G

Cuando se habla de una nueva red móvil 5G también se deben considerar los cambios que estos sistemas van a aportar, no sólo mejorarán las comunicaciones móviles de banda ancha, sino que también permitirán hacer grande y diversa la aplicación de esta tecnología a casos de utilización en los que estén implicadas comunicaciones ultra fiables y de baja latencia, y comunicaciones masivas entre máquinas (ITU, 2017).

Para una mejor explicación de los casos de uso con más relevancia para las futuras redes de comunicaciones móviles, haciendo notar sus desafíos y requisitos se puede considerar el 5G e IoT ya que serán de suma importancia para proveer todas las características técnicas necesarias a cada una de estas nuevas tendencias. Es importante mencionar que ni el 5G ni las tecnologías para IoT deberán atender en su totalidad todos los casos de uso, sino que, para eso, debe haber también una cooperación entre los sistemas. En la Figura 2 se muestran los sectores socio económicos que se relacionan con algunos casos de uso atendidos por los sistemas 5G e IoT, como se detalla a continuación: la Agricultura, se relaciona con la automatización; la Construcción, se relaciona con la automatización y con Smart Cities[19]; la Energía, se relaciona con la automatización, Smart Cities y teleprotección en Smart Grid[20]; las Finanzas, se relacionan con Smart Cities, y con los dispositivos electrónicos que se

[17] UDM: Unified Data Management
[18] PM: Performance Management
[19] Smart Cities: Ciudades Inteligentes.
[20] Smart Grid: Red de distribución de recursos.

usan en el cuerpo humano o también llamados Wearables; en el caso de la Manufactura, se relaciona con la automatización, la realidad virtual, la realidad aumentada y teleprotección en Smart Grid; en cuanto a los Medios, se relacionan con procedimientos médicos, realidad virtual, realidad aumentada y con wearables; también se hace referencia a un sector socio económico muy importante como es la Salud, que se relaciona con procedimientos médicos y con wearables; la Seguridad Pública es otro sector socioeconómico considerable y se relaciona con la automatización, procedimientos médicos, Smart Cities y vehículos; finalmente se encuentra el Transporte que se relaciona con la automatización, Smart Cities, y vehículos autónomos; como se puede ver existen algunos casos de uso que se relacionan con uno o varios sectores socio económicos (Magalhães, n.d.).

Figura 2 *Sectores socioeconómicos y casos de uso.*

Casos de uso / Sectores socioeconómicos	Automatización	Procedimientos médicos	Realidad Virtual y Realidad Aumentada	Smart Cities	Teleprotección en Smart Grid	Vehículos Autónomos	Wearables
Agricultua	■						
Construcción	■			■			
Energía	■			■	■		
Finanzas				■			■
Manufactura	■		■		■		
Medios		■	■				■
Salud		■					■
Seguridad Pública	■	■		■		■	
Transporte	■			■		■	

Fuente: (Magalhães, n.d.)

A continuación, se realiza una breve descripción de los casos de uso de 5G e IoT para una mejor comprensión.

Automatización Industrial

Con la llegada de la industria 4.0, nuevas tecnologías de automatización industrial demandan la comunicación a distancia y que a su vez esta sea segura y con rápido tiempo de respuesta para que haya un control y monitoreo del proceso productivo, además se requiere la comunicación entre máquinas y esto puede ser un factor determinante para facilitar y optimizar los procesos.

Aunque actualmente estos procesos no demanden de la red la transferencia de grandes volúmenes de datos, y a su vez no necesiten de soporte de alta movilidad, es necesario que se cuente con una red con ciertas características como baja latencia y alta confiabilidad. Debido a esto, actualmente este nivel de automatización es hecho por redes fijas, sin embargo, este puede ser un factor limitante para la evolución de la industria en áreas remotas, y por lo tanto, se ve la necesidad de la implementación conjunta con redes IoT y 5G. (Magalhães, n.d.)

Procedimientos Médicos Remotos

Las tecnologías médicas a nivel mundial también han tenido grandes avances porque esto ha permitido que sean realizados procedimientos desde los considerados simples hasta incluso intervenciones quirúrgicas en pacientes, para ello utilizan instrumentos de alta tecnología o incluso robótica. Estos procedimientos son esenciales para la vida ya que demandan de una gran experiencia profesional y del instrumental específico, los procedimientos pueden ser realizados remotamente a través de la red 5G, en donde, un profesional especializado podrá operar pacientes en cualquier lugar del mundo, manipulando instrumentos robóticos y orientándose por imágenes de alta resolución. Sin duda las cirugías remotas ofrecen oportunidades para que pacientes que se encuentran en áreas aisladas reciban servicios de salud a tiempo y con costos accesibles. Para que esto sea posible el uso de estos servicios la red necesita debe cumplir con comunicaciones ultra confiables y con baja latencia punta a punta, para que se pueda tener la sensación de interacción táctil (movimientos con respuesta prácticamente instantánea). Otro punto importante a considerar en cuanto a la medicina es el soporte a vehículos de rescate (ambulancias) y la conectividad con dispositivos IoT de monitoreo y soporte, sin duda este es un gran aporte ya que mejorará en gran manera la entrega de informaciones y comunicación con hospitales y afines. (Magalhães, n.d.).

Realidad Virtual y Realidad Aumentada

Estas tecnologías proporcionan a los usuarios la capacidad de interacción, unos con otros, como si estuviesen físicamente en un mismo lugar. Al contar con estas capacidades se proporciona a los usuarios algunas posibilidades de interacción, entre ellas se tienen: conferencias, reuniones, juegos y reproducción de medios de una manera que antes no ha sido experimentada, además es importante mencionar que se hace posible que personas con habilidades específicas localizadas remotamente puedan realizar tareas complicadas en conjunto. Para poder crear la sensación de inmersión, los usuarios de estas tecnologías deben permanecer constantemente actualizados vía streaming de datos, unos con otros, ya que todos los miembros pueden afectar los escenarios creados por estas tecnologías. Para que se pueda tener una buena experiencia, es indispensable que una cantidad significativa de información sea intercambiada a tiempo, en ambas direcciones, entre sensores y dispositivos de los usuarios y la nube, es decir que los flujos multidireccionales se deben manejar con altas tasas de transferencia de datos y bajísima latencia ya que son necesarios para mantener alta calidad de servicio (Magalhães, n.d.).

Smart Cities

Este es un conceto que cada vez se hace más conocido y hace referencia a que la conectividad que actualmente es proporcionada en su mayoría entre personas, posteriormente será utilizada también para conectar a las personas con el ambiente que se encuentra a su alrededor. Cuando se habla de Smart Cities no se limita sólo a ciudades ya que puede ser extendido también a conceptos como casas inteligentes, edificios inteligentes, oficinas inteligentes y afines. Estos escenarios consisten en que pueda tener la coexistencia de una serie de dispositivos con una variedad diversa de tipos y funcionalidades, trabajando en conjunto para proporcionar un ambiente que sea adaptable, activo, que sea seguro y que permita realizar el monitoreo y configuraciones remotas. Para que pueda llevarse a cabo esta diversidad de servicios, los requisitos de red como se ha dicho anteriormente también deben ser mejorados. Por ejemplo, cuando se habla de la transferencia de archivos a la nube, este proceso demanda de altas tasas de datos, pero cuando se habla de pequeños dispositivos para pagos, wearables, sensores y actuadores, en su mayoría no se requiere de altas tasas, pero por otro lado si se requiere de baja latencia, también depende de la concentración espontánea de personas, que se encuentren en esos escenarios ya que se pueden encontrar en ambientes abiertos o cerrados (Magalhães, n.d.).

Teleprotección en Smart Grid

Las redes Smart Grid consisten en una red de distribución de recursos (electricidad, agua, gas) estas redes se caracterizan porque utilizan las tecnologías de la información para hacer el sistema más eficiente, confiable y sustentable. Al utilizar la red móvil de quinta generación para administrar estos sistemas, se requiere contar con la capacidad de reacción rápida a cambios en la entrega o uso de los recursos, con esto se evitarían fallas que interrumpan y afecten el servicio o impacten de forma crítica a la sociedad. Un claro ejemplo sería una un apagón, que puede ser consecuencia de un daño causado por un evento imprevisto. En este caso, el monitoreo y control de los sistemas en conjunto con soluciones inalámbricas tienen un papel vital en la identificación y posterior corrección del problema, ya que se genera el intercambio de información crítica de manera confiable, esto tiene mucha utilidad para que haya reacción rápida a fin de evitar daños críticos antes de problemas como ese, los sistemas de teleprotección exigen baja latencia y alta confiabilidad en la transferencia de informaciones. (Magalhães, n.d.)

Vehículos Autónomos

Cuando se habla de conducción autónoma se hace referencia a un destino inevitable para el futuro de los automóviles, que sin duda proporcionará no solo confort y comodidad a las personas, sino que también contribuirá a reducir drásticamente el número de imprudencias y accidentes de tránsito. Para que sea posible la conducción autónoma, se requiere no solo la comunicación de los vehículos con la infraestructura de red, sino que también debe existir la comunicación de los vehículos entre si (vehículo a vehículo), del vehículo con el conductor y del vehículo con sensores u otros dispositivos. Esta conectividad necesita tener una baja latencia y sin duda alta confiabilidad para que haya seguridad en el control de los vehículos. Aunque este tipo de señalización no exija la transferencia de grandes volúmenes de datos, algunas aplicaciones, si lo pueden requerir como, por ejemplo, intercambio de información de video entre vehículos con el fin de que se pueda llevar a cabo el control y monitoreo de por ejemplo de una flota, esto puede exigir de la red tasas de datos más elevadas, por ello también se ve la necesidad de contar con el soporte de alta movilidad, ya que estos vehículos pueden, por ejemplo, atravesar por células 5G durante su trayecto. (Magalhães, n.d.)

Wearables

Este es un término que se usa para describir a las "tecnologías vestibles", es decir, son los dispositivos que pueden ser usados en el cuerpo humano como piezas de vestuario o accesorios, y, así, traer varios beneficios a los usuarios, como, por ejemplo, se puede realizar el monitoreo de funciones corporales y de actividades físicas, o también obtener características estéticas. La mayoría de estos dispositivos no necesitan tener una conexión constante con la red, como es el caso, por ejemplo, de dispositivos que son solamente estéticos, existen otros que sí demandan cierto nivel de conectividad, como Smartwatches[21], Smartbands[22], rastreadores para ancianos, sensores para funciones específicas, entre otros. En este último caso y teniendo en cuenta que se puede llegar a tener una alta cantidad de dispositivos si se requiere que la red tenga, disponibilidad, latencia relativamente baja y confiabilidad. (Magalhães, n.d.)

Relación de la Tecnología 5G y Big Data

Las redes móviles actuales no manejan la cantidad masiva de datos que se prevé tener en la red móvil 5G, es por ello que a medida que las redes móviles evolucionan y como se ha mencionado anteriormente, la tecnología 5G se involucra en todos los aspectos de la sociedad, es importante considerar que los beneficios que esta tecnología promete al trabajar a altas velocidades, así como también al tener un notable aumento de la densidad máxima de conexiones y tener una transmisión masiva y constante de datos, requiere la potencialidad que ofrece el Big Data, esta tecnología a través de su arquitectura permite la ingestión y tratamiento de datos de forma masiva como se requiere en la tecnología 5G, es necesaria la combinación de estas dos tecnologías, para que sea posible tener los escenarios deseados. Al trabajar conjuntamente 5G y Big Data, harán posible la conexión de innumerables dispositivos nuevos. Big Data no solo permite la ingestión de datos sino también su análisis para luego brindar servicios, da una gran cantidad de oportunidades a los operadores de redes móviles para mejorar la calidad de servicio y la calidad de experiencia del usuario ya que a través de

[21] Smartwatches: Son relojes inteligentes.

[22] Smartbands: Son pulseras inteligentes.

las técnicas de recopilación de grandes cantidades de datos y análisis de los mismos se tiene una optimización de la red. (Zheng et al., 2016)

Big Data

Son los conjuntos de datos con gran volumen y estructura que excede las capacidades de las herramientas de programación tradicionales (bases de datos, software, etc.), para la recopilación, almacenamiento y procesamiento de datos en un tiempo razonable y excede la capacidad de percepción por un humano. Los datos que se manejan en Big Data pueden ser: estructurados, semiestructurados y no estructurados, esto hace que sea imposible gestionarlos y procesarlos de manera efectiva de manera tradicional. (Miloslavskaya & Tolstoy, 2016)

Uno de los grandes aportes de esta tecnología y que se considera una de las claves del auge del Big Data consiste en que esta ofrece la posibilidad de gestionar no solo datos estructurados, como venían haciendo las bases de datos tradicionales, sino también es capaz de gestionar datos no estructurados, además de que se puede gestionar una mayor cantidad de datos en menos tiempo y también cuenta con la capacidad de procesar información de distintos formatos, tipos o frecuencia en la que se generan (Iñigo, 2020).

Arquitectura Big Data

Como ya se ha mencionado Big Data tiene la capacidad de gestionar una cantidad masiva de datos sin la necesidad de que sean estructurados, esto es posible gracias a la arquitectura de cinco capas que forman cualquier proyecto de Big Data. Las cinco capas son las siguientes: fuentes de datos, recolección de datos, almacenamiento, análisis y procesamiento. (Iñigo, 2020)

La arquitectura referencial de Big Data para la gestión de las telecomunicaciones sería la arquitectura Lambda ya que está se caracteriza por utilizar distintas capas para el procesamiento batch y el streaming, esto es beneficioso para la red móvil porque permite por ejemplo que la capa batch pueda encargarse de entrenar modelos e ir mejorando las predicciones en base a los datos que se reciben de

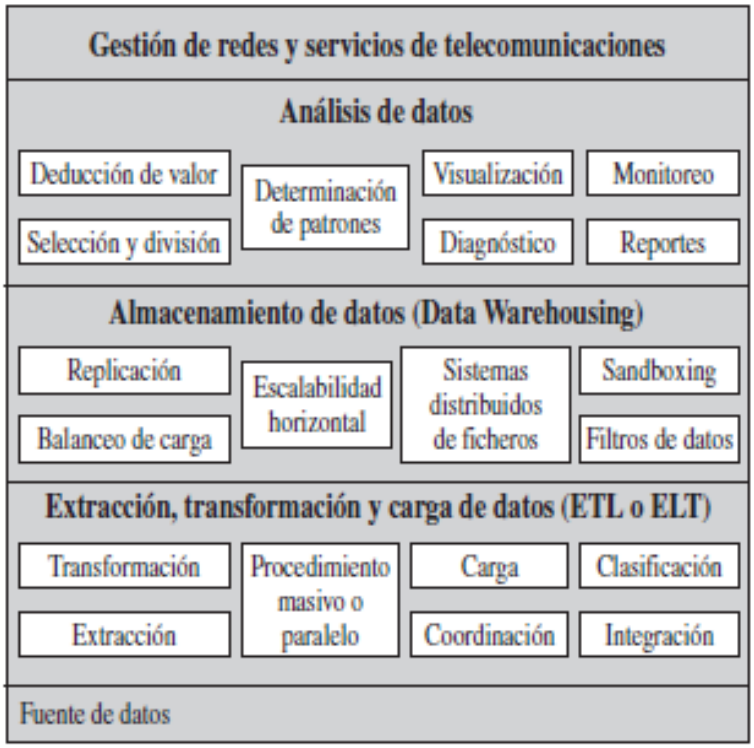
los usuarios; y que por otro lado la capa streaming se encargue de las valoraciones en tiempo real. Es una arquitectura de procesamiento genérica para Big Data, sin embargo, posee algunas características que la han hecho ser una de las arquitecturas mayormente implementadas cuando se busca procesar información de grandes volúmenes como se menciona en (Careaga, 2017). Además que la arquitectura Lambda está orientada a aplicaciones que han sido construidas alrededor de transformaciones asíncronas complejas que necesitan ejecutarse con latencia baja (segundos a minutos) como es el caso también de la red móvil 5G (Algorri Álvarez, 2017).

Como se puede apreciar en la Figura 3, en el nivel más bajo de la arquitectura, se encuentran las fuentes que generan grandes flujos de datos a diferentes velocidades y desde distintos puntos geográficos. En un segundo nivel, aparecen los procesos de extracción, transformación y carga, el objetivo es extraer los datos de distintas fuentes y enviarlos a los repositorios donde se almacenan. Los procesos de transformación y carga se pueden dar de dos formas principales: la primera, los datos inicialmente son cargados en las bases de datos que los almacenarán y dentro de estas se hacen las transformaciones necesarias, lo que facilita que las herramientas de análisis de datos los procesen y entreguen información clara y entendible, y; la segunda, en donde los datos son transformados previamente al almacenamiento de los mismos. En el tercer nivel de la arquitectura se considera el almacenamiento de datos masivos. Este nivel puede variar de una implementación a otra de la arquitectura, puesto que existen herramientas ETL que no solo transforman los datos, sino que presentan espacios de almacenamiento para grandes volúmenes de información, no requiriéndose emplear bases de datos adicionales. Es importante mencionar que cada empresa u organización donde se aplique la arquitectura que se propone, puede determinar, de acuerdo a los tipos de datos con los que trabaje, cómo almacenarlos (Plasencia Moreno & Anías Calderón, 2017).

En el cuarto nivel de la arquitectura se considera el análisis de datos, en el que se emplean herramientas que se encargarán de obtener información de alto nivel de impacto, que sea útil para la gestión de las redes y servicios de telecomunicaciones. En este nivel se emplean herramientas de análisis predictivo de datos, algoritmos que establecen puntos de interrelación dentro de grandes volúmenes de datos, herramientas de visualización que permitan representar información de interés sobre las redes y servicios para una empresa de telecomunicaciones (Plasencia Moreno & Anías Calderón, 2017).

Finalmente, en el último nivel de la arquitectura referencial propuesta, se encuentra la gestión de las redes y servicios de telecomunicaciones, esta se puede optimizar gracias al análisis de los datos masivos para, por ejemplo, lograr la configuración eficiente de los dispositivos de interconexión de redes, la mejora en los servicios telefónicos, una mayor calidad de las ofertas a los clientes y la optimización de las redes (Plasencia Moreno & Anías Calderón, 2017).

Figura 3 *Propuesta de arquitectura referencial de Big Data para la gestión de las telecomunicaciones.*



Fuente: (Plasencia Moreno & Anías Calderón, 2017)

Ingestión

Por Ingestión se entiende al proceso de introducción de datos en el sistema que se está construyendo o utilizando, los procedimientos de procesamiento de Big Data normalmente comienzan por recopilar un gran volumen de datos de múltiples fuentes y es por ello que es de gran importancia el proceso

que se realiza en la capa de Ingestión. Primero, los datos se recopilan en fragmentos en varios caminos para una utilización eficiente de la red y confiable entrega. En segundo lugar, los datos se ingieren constantemente, tercero, los datos se transfieren a través de Internet. (Yoon & Kim, 2018)

Sin embargo, la complejidad de llevar a cabo la ingestión de datos depende de la calidad de los datos y de cuán lejos estén del formato deseado antes de iniciar el procesamiento, una de las maneras para agregarlos es mediante las herramientas de ingestión proporcionadas por los entornos de trabajo. (Sánchez, 2019)

Durante los últimos años, la tecnología tuvo un gran impacto en las aplicaciones y en el procesamiento de datos, y las organizaciones han comenzado a darle más importancia a los datos e invertir más en su recolección y gestión. El Big Data creó además una nueva era y nuevas tecnologías que permiten analizar tipos de datos como texto y voz, que tienen un enorme volumen en Internet y en otras estructuras digitales. La evolución de los datos es muy importante mencionar ya que, en el pasado, el volumen de datos estaba al nivel de los bytes y hoy en día las empresas utilizan un volumen enorme de datos al nivel de los petabytes. Los expertos del Centro Nacional de Datos Climáticos en Asheville estimaron que si se desea almacenar todos los datos que existen en el mundo, se necesita al menos 1200 exabytes, pero es imposible precisar un número relevante (Matacuta & Popa, 2018).

En muchas situaciones, cuando se utiliza Big Data, se desconoce la fuente de la estructura de los datos y si en las empresas, por ejemplo, se utilizan los métodos comunes de ingestión de datos, es difícil que se puedan manipular los datos. Para las empresas la ingesta de datos representa una estrategia importante, ya que esto representa una gran ayuda ya que les ayuda a retener clientes y obtener una mayor rentabilidad. (Matacuta & Popa, 2018)

Herramientas de ingestión de datos

Son herramientas que permiten realizar: la extracción o recolección de datos desde la fuente, luego está la transformación, es decir, validar, limpiar y normalizar los datos asegurándose de su precisión y confiabilidad; y finalmente la carga, donde se colocan los datos en el repositorio o base de datos correcta para su análisis posterior (Maldonado, 2018).

Data Lake

Un lago de datos se refiere a un repositorio de almacenamiento masivamente escalable que contiene una gran cantidad de datos sin procesar en su formato nativo ("tal cual") o también llamados datos en bruto. Los lagos de datos generalmente se construyen para manejar grandes volúmenes de datos no estructurados que llegan rápidamente. Por esta razón, los datos en el repositorio se vuelven accesibles y a menudo incluyen una base de datos semántica, un modelo conceptual que aprovecha el mismo, estándares y tecnologías utilizados para crear hipervínculos de Internet y agregar una capa de contexto sobre los datos (Miloslavskaya & Tolstoy, 2016)

Para probar las herramientas de ingestión de datos y enviarlas al Datalake se requieren archivos de prueba. En la red móvil 5G se utilizan los archivos de gestión del rendimiento que como ya se mencionó anteriormente sirven para monitorear y medir continuamente el desempeño de la red, estos datos de PM, son de mucha utilidad ya que se puede planificar y optimizar la red para evitar cortes y brindar una mejor experiencia a los usuarios, como se menciona en la investigación realizada por (Martinez et al., 2020) el formato utilizado para archivos PM es XML en tecnologías 2G, 3G, 4G y 5G para la mayoría de los proveedores, por lo tanto, los archivos que se utilizan para las pruebas de ingestión de datos tienen el formato xml, mismos que contienen información sobre las mediciones de una red de prueba, es decir no son datos generados por el usuario.

Archivo XML

Para realizar las pruebas correspondientes con las herramientas de ingestión de datos se requieren archivos 5G con formato XML, el mismo que se detalla en el anexo A, también es importante mencionar que los archivos XML tienen etiquetas que aportan información y sirven para describir los datos y la estructura de los mismos, a continuación, en la tabla 1 se explican las etiquetas que 3GPP ha definido como especificaciones técnicas y aspectos del sistema para los archivos XML en la gestión de telecomunicaciones, las cuales se han mantenido lo más cortas posible para minimizar el tamaño de los archivos de resultados de medición XML[23].

[23] XML: Extensible Markup Language

Tabla 1 *Descripción de parámetros de las etiquetas XML*

ASN.1 Tag	XML tag	Description
MeasDataCollection	mdc	Esta es la etiqueta de nivel superior, que identifica el archivo como una colección de datos de medición. El contenido del archivo se compone de un encabezado ("MeasFileHeader"), la colección de elementos de resultado de medición ("MeasData") y un pie de página del archivo de medición ("MeasFileFooter").
measFileHeader	mfh	Este es el encabezado del archivo de resultados de medición que se insertará en cada archivo. Incluye un indicador de versión, el nombre, tipo y nombre del proveedor del nodo de red emisor, y una marca de tiempo ("collectionBeginTime").
measData	md	La construcción "MeasData" representa la secuencia de cero o más elementos de resultado de medición contenidos en el archivo. Puede estar vacío en caso de que no se puedan proporcionar datos de medición. Los elementos individuales "MeasData" pueden aparecer en cualquier orden. Cada elemento "MeasData" contiene el nombre del NE ("nEId") y la lista de resultados de medición pertenecientes a ese NE ("MeasInfo").
measFileFooter	mff	El pie de página del archivo de resultados de medición que se insertará en cada archivo. Incluye una marca de tiempo, que se refiere al final del intervalo de recopilación de medición general que cubre los resultados de medición recopilados que se almacenan en este archivo.

fileFormatVersion	ffv	Este parámetro identifica la versión de formato de archivo aplicada por el remitente. La versión de formato definida en el presente documento será "2" para los formatos XML y ASN.1 por igual.
senderName	Sn	SenderName identifica de forma única el NE o EM que ensambló este archivo de medición por su nombre distinguido (DN), de acuerdo con las definiciones en 3GPP TS 32.300 [10]. En el caso del enfoque basado en NE, es idéntico al "nEDistinguishedName" del remitente. La cadena puede estar vacía (es decir, tamaño de cadena = 0) en caso de que el DN no esté configurado en el remitente.
senderType	St	Este es un identificador configurable por el usuario del tipo de nodo de red que generó el archivo, p. NodoB, EM, SGSN. La cadena puede estar vacía (es decir, tamaño de cadena = 0) en caso de que "senderType" no esté configurado en el remitente.
vendorName	vn	El "vendorName" identifica al proveedor del equipo que proporcionó el archivo de medición. La cadena puede estar vacía (es decir, tamaño de cadena = 0) si el "vendorName" no está configurado en el remitente.
collectionBeginTime	cbt	"CollectionBeginTime" es una marca de tiempo que se refiere al inicio del primer intervalo de recopilación de mediciones (período de granularidad) que está cubierto por los resultados de medición recopilados que se almacenan en este archivo.
nEId	neid	La identificación única del NE en el sistema. Incluye el nombre de usuario ("nEUserName"), el

		nombre distinguido ("nEDistinguishedName") y la versión de software ("nESoftwareVersion") del NE.
nEUserName	neun	Este es el nombre definible por el usuario ("userLabel") definido para el NE en 3GPP TS 32.622 [24]. La cadena puede estar vacía (es decir, tamaño de cadena = 0) si "nEUserName" no está configurado en las aplicaciones CM.
nEDistinguishedName	nedn	Este es el nombre distinguido (DN) definido para el NE en 3GPP TS 32.300 [10]. Es único en la red 3G de un operador. La cadena puede estar vacía (es decir, tamaño de cadena = 0) si "nEDistinguishedName" no está configurado en las aplicaciones CM.
nESoftwareVersion	nesw	Esta es la versión de software ("swVersion") definida para el NE en 3GPP TS 32.622 [24]. Este es un parámetro opcional que permite que los sistemas de posprocesamiento se encarguen de las mediciones específicas del proveedor modificadas entre versiones de software.
measInfo	mi	Esta es la etiqueta de nivel superior, que identifica el archivo como una colección de datos de medición. El contenido del archivo se compone de un encabezado ("MeasFileHeader"), la colección de elementos de resultado de medición ("MeasData") y un pie de página del archivo de medición ("MeasFileFooter").
measTimeStamp	mts	Marca de tiempo que se refiere al final del período de granularidad.
granularityPeriod	gp	Período de granularidad de la (s) medición (es) en segundos.

measTypes	mt	Esta es la lista de tipos de medición a la que pertenece la siguiente lista análoga de valores de medición ("MeasValues"). Los tipos de medición solo GSM se definen en TS 52.402 [22]. Los tipos de medición para implementaciones UMTS y UMTS / GSM combinados se especifican en TS 32.403 [23].
measValues	mv	Este parámetro contiene la lista de resultados de medición para el recurso que se mide, p. Ej. tronco, celda. Incluye un identificador del recurso ("MeasObjInstId"), la lista de valores de resultado de medición ("MeasResults") y una bandera que indica si los datos son confiables ("Sospechoso Bandera").
measObjInstId	moid	El campo "MeasObjInstId" contiene el nombre distinguido local (LDN) del objeto medido dentro del alcance definido por "nEDistinguishedName" (ver 3GPP TS 32.300 [10]). La concatenación de "nEDistinguishedName" y "MeasObjInstId" produce el DN del objeto medido. Por tanto, el "MeasObjInstId" está vacío si "nEDistinguishedName" ya especifica completamente el DN del objeto medido, que es el caso de todas las mediciones especificadas en el nivel NE. Por ejemplo, si el objeto medido es un "ManagedElement" que representa RNC "RNC-Gbg-1", entonces el "nEDistinguishedName" será, por ejemplo, "DC = a1.companyNN.com, SubNetwork = 1, IRPAgent = 1, SubNetwork = CountryNN, MeContext = MEC-Gbg-1, ManagedElement = RNC-Gbg-1 "y el"

		MeasObjInstId "estará vacío. Por otro lado, si el objeto medido es una "UtranCell" que representa la celda "Gbg-997" administrada por ese RNC, entonces el "nEDistinguishedName" será, por ejemplo, el mismo que el anterior, es decir, "DC = a1.companyNN.com, SubNetwork = 1, IRPAgent = 1, SubNetwork = CountryNN, MeContext = MEC-Gbg-1, ManagedElement = RNC-Gbg-1 "y el" MeasObjInstId "será, por ejemplo," RncFunction = RF-1, UtranCell = Gbg-997 ". La clase de "MeasObjInstId" se define en el elemento F de cada plantilla de definición de medición.
measResults	R	Este parámetro contiene la secuencia de valores de resultado para los tipos de medición observados. La secuencia "MeasResults" debe tener el mismo número de elementos, que siguen el mismo orden que la secuencia MeasTypes. Los valores normales son INTEGER y REAL. El valor NULL está reservado para indicar que el elemento de medición no es aplicable o no se pudo recuperar para la instancia del objeto.
suspectFlag	sf	Se utiliza como indicación de la calidad de los datos escaneados. FALSO en el caso de datos confiables, VERDADERO si no es confiable. El valor predeterminado es "FALSO", en caso de que la bandera sospechosa tenga su valor predeterminado, puede omitirse.
TimeStamp	ts	Formato ASN.1 GeneralizedTime. La información mínima requerida dentro de la marca de tiempo es año, mes, día, hora, minuto y segundo.

Fuente:(3rd Generation Partnership Project, 2020)

Marco Legal

Existen Instituciones para la Regulación y Control de las redes móviles a nivel nacional como el Ministerio de Telecomunicaciones (MINTEL) y la Agencia de Regulación y Control de las Telecomunicaciones (ARCOTEL) y a nivel internacional están la Unión Internacional de Telecomunicaciones, Facilitadores de comunicaciones móviles e inalámbricas para la sociedad de la información veinte (METIS por sus siglas en inglés), 5G LAB y el Proyecto Asociación de Tercera Generación (3GPP por sus siglas en inglés), a continuación se hará un detalle de cada una.

MINTEL

Aunque no se cuenta con un estándar de 5G en Ecuador, el Ministerio de Telecomunicaciones y de la Sociedad de la Información, en su programa de acción ECUADOR CONECTADO promueve fomentar la licitación de espectro para la masificación de 4G y despliegue de 5G, impulsando tecnologías emergentes como Internet de las cosas y Big Data. (Ministerio de Telecomunicaciones y de la Sociedad de la Información, 2019).

ARCOTEL

La Agencia de Regulación y Control de las Telecomunicaciones es la entidad encargada de la administración, regulación y control de las telecomunicaciones y del espectro radioeléctrico y su gestión, así como de los aspectos técnicos de la gestión de medios de comunicación social que usen frecuencias del espectro radioeléctrico o que instalen y operen redes. (ARCOTEL, 2020)

UIT

La Unión Internacional de Telecomunicaciones (UIT), impulsa la creación de estándares y prototipos por parte de los entes, organizaciones e instituciones por medio del intercambio de los resultados obtenidos en sus respectivos proyectos. Para finalizar, desde el año 2020 en adelante,

se tiene previsto las pruebas y desarrollo comercial de la tecnología 5G. (González & Salamanca, 2016)

Es importante mencionar que hasta el año 2018 no se han establecido totalmente las especificaciones de regulación y los estatutos en base a los cuales se cimentará 5G a nivel mundial, pero desde el año 2015 la UIT bajo el estándar de IMT-2020 se pueden estudiar las características de interacción de las redes 5G en el futuro y cuál será el impacto de estas.(Vera Cárdenas, 2018)

METIS

La perspectiva de METIS, el cual es un consorcio de 29 socios coordinados por Ericsson, que incluye fabricantes, operadores de telecomunicaciones, instituciones académicas, industria automotriz y un centro de investigación, donde el objetivo es sentar las bases de 5G, el sistema de comunicaciones móviles e inalámbricas de próxima generación. (González & Salamanca, 2016)

5G LAB

El 5G Lab de Alemania es un consorcio de tecnología reconocido en el esfuerzo colaborativo requerido para desarrollar y repartir redes de 5G. Dicho laboratorio comprende cuatro rutas diferentes que permite a los miembros enfocarse en áreas de interés, a su vez previendo una visión general de las redes 5G. (González & Salamanca, 2016)

3GPP

Es el organismo de estandarización que especifica los sistemas de comunicaciones móviles, está establecido por una colaboración de varios grupos de asociaciones de telecomunicaciones, se les conoce como miembros organizativos, estos miembros del Proyecto de Asociación de Tercera Generación, se reúnen regularmente para colaborar y crear estándares de comunicaciones celulares. Los estándares 3GPP están estructurados en releases, actualmente, 3GPP está definiendo estándares para 5G. (Díaz Zayas et al., 2017).

CAPITULO III: MARCO METODOLÓGICO

Descripción del área de estudio

Una potencial área de estudio de este trabajo de investigación son las empresas ecuatorianas y extranjeras las cuales se enfoquen en la recolección de datos de forma masiva y que su giro de negocio tenga que ver con la red móvil y que quieran evaluar su crecimiento, mejorar sus servicios y tecnologías, tales como: CNT, MOVISTAR, CLARO.

Enfoque y tipo de investigación

Según el objeto de estudio esta investigación es principalmente del tipo aplicada, de acuerdo con el nivel de medición y análisis de información se puede categorizar con enfoque cuantitativo ya que las pruebas se basan en el tiempo que se demoran en el procesamiento de ingestión cada una de las herramientas.

Procedimiento de investigación

Figura 4 *Proceso de investigación a seguir en este trabajo.*

Nota: Realizado por el Autor

- Revisión de literatura en el marco teórico sobre: 5G, Big Data, ingestión, Data Lake y aplicaciones de recolección de datos.
- Definición del formato de datos a usarse en una red móvil 5G en el marco teórico de esta investigación.
- Comparación de diferentes herramientas de recolección de datos para usar las que cumplan con los requerimientos.
- Desarrollo de pruebas con las herramientas seleccionadas.
- Análisis de resultados.
- Recomendación de la herramienta de recolección de datos más eficiente en base al tiempo de procesamiento, consumo de CPU y consumo de memoria RAM.

Selección de Herramientas

Se han tomado en cuenta algunas investigaciones de las herramientas más recomendadas en (Matacuta & Popa, 2018), (Cacho, 2014), (Fernández, 2017) y (Sánchez, 2019) y se ha examinado las herramientas de Apache porque Apache es muy conocido en el área de desarrolladores y es el software de servidor web más utilizado, que se ejecuta en el 67% de los servidores web de todo el mundo (Matacuta & Popa, 2018). A continuación, se presenta una breve descripción de las herramientas de ingestión de datos.

Apache Nifi: Es un sistema de gestión de flujo de datos que viene con una interfaz de usuario web que ayuda a construir flujos de datos en tiempo real, admite programación basada en flujo y la programación gráfica incluye procesadores y conectores, en lugar de nodos y bordes". Una característica importante es la capacidad de Nifi de ingerir cualquier dato utilizando metodologías de ingestión para cualquier dato en particular. Similar a ingresar datos. (Matacuta & Popa, 2018)

Apache Sqoop: Esta herramienta permite la transferencia bidireccional de datos entre Hadoop y una base de datos SQL (datos estructurados).

Apache Flume: Es un servicio distribuido, confiable y disponible para recopilar, agregar y mover de manera eficiente grandes cantidades de datos de registro ". es una definición confiable que se encuentra en el sitio web oficial de Apache Flume y es un software listo para producción. Esta herramienta está diseñada para ingerir y recopilar grandes volúmenes de datos de múltiples fuentes en Hadoop Distributed File System (HDFS).

Apache Kafka: Es un sistema de almacenamiento publicador/subscriptor distribuido, particionado y replicado. Estas características, añadidas a que es muy rápido en lecturas y escrituras lo convierten en una herramienta excelente para comunicar streams de información que se generan a gran velocidad y que deben ser gestionados por una o varias aplicaciones, para comprender su funcionamiento es importante examinar sus componentes (Moraga, 2018).

El Kafka Clúster, es una colección de uno o más servidores conocidos como brokers; el productor, es el componente que se utiliza para publicar los mensajes; el consumidor, es el componente que se utiliza para recuperar o consumir mensajes; para las pruebas realizadas se ha utilizado un clúster de Kafka. (Moraga, 2018)

Para que Kafka pueda funcionar correctamente requiere la instalación de ZooKeeper que es un servicio de coordinación centralizado que se utiliza para mantener la información de configuración en los nodos del clúster en un entorno distribuido. (Toasa, 2015)

La unidad fundamental de datos en Kafka es un mensaje. Kafka convierte todos los mensajes en matrices de bytes. Cabe señalar que las comunicaciones entre los productores, consumidores y clústeres en Kafka utilizan el protocolo TCP. Cada servidor o broker de un clúster de Kafka se conoce como intermediario. Puede escalar Kafka horizontalmente simplemente agregando agentes adicionales al clúster (Moraga, 2018).

Kafka Connect es una herramienta para transferir datos de manera fiable y escalable entre Apache Kafka y otros sistemas. Ofrece la posibilidad de definir conectores de forma rápida y sencilla que son capaces de mover grandes cantidades de datos hacia Kafka o de Kafka hacia afuera, En los conectores de Kafka se tienen los workers, que son instancias en las cuales se realiza un procedimiento de conexión, conversión y serialización (Moraga, 2018).

Apache Spark: Es un sistema de computación distribuida de software libre, que permite procesar grandes conjuntos de datos sobre un conjunto de máquinas de forma simultánea, proporcionando escalabilidad horizontal y la tolerancia a fallos.

Para cumplir con estas características proporciona un modelo de desarrollo de programas que permite ejecutar código de forma distribuida de tal manera que cada máquina se ocupe de realizar una parte de la tarea y entre todos realicen la tarea global.

Utiliza estructuras de datos resilientes distribuidos RDD (RDD por sus siglas en inglés) que son un conjunto de datos de solo lectura, y que están distribuidos a lo largo del clúster, mantenidos de manera tolerante a fallos. La disponibilidad de RDDs facilita la implementación de algoritmos iterativos que accedan varias veces a los mismos datos y para el análisis exploratorio de datos.

Para su correcto funcionamiento Spark necesita:

- Gestión de recursos, soporta Standalone, YARN o Apache Mesos.
- Sistema de ficheros distribuido, soporta HDFS, Cassandra o Kudu.

RabbitMQ: Es un sistema colas de mensajes que actúa de middleware entre productores y consumidores.

Amazon Kinesis: Es el homólogo de Kafka para la infraestructura Amazon Web Services.

Microsoft Azure Event Hubs: Es el homólogo de Kafka para la infraestructura Microsoft Azure.

Hadoop: El ecosistema de Apache Hadoop es una estructura de soporte esencial para procesar y almacenar una gran cantidad de datos. Este ecosistema de Apache Hadoop crece continuamente y consta de múltiples proyectos y herramientas con características y beneficios valiosos que brindan capacidad de carga, transferencia, transmisión, indexación, mensajería, consultas y muchos otros. Hadoop contiene dos elementos principales: Hadoop Distributed Filesystem (HDFS) y MapReduce. El HDFS es un sistema de archivos diseñado para el almacenamiento y procesamiento de datos. Hay muchos subproyectos (administrados principalmente por Apache, que son realizados por organizaciones libres) diseñados para el mantenimiento y el monitoreo que se integran muy bien con Hadoop y nos permiten concentrarnos en desarrollar la ingestión de datos en lugar de monitorearlos.

El sistema de distribución de archivos de Hadoop, soporta listas de control de acceso y el modelo tradicional de permisos de acceso a archivos. Para el control de usuarios y lanzamiento de procesos, Hadoop cuenta con servicio de nivel de autorización, que asegura que los clientes tengan los permisos correctos (Salas, 2017).

Para realizar la selección de herramientas de ingestión de datos, se ha utilizado la norma IEEE 29148, esta sirve para detallar el tratamiento unificado de los procesos y productos que se involucran en la gestión de requerimientos durante todo el ciclo de vida de los sistemas y el software, además también proporciona detalles para la construcción de requisitos textuales bien formados, que sirven para incluir características y atributos, en el contexto de la ingeniería de sistemas y software, se menciona también que la calidad es un punto importante a considerar ya que es el conjunto de atributos que tiene un software para satisfacer los requerimientos expresados por los interesados (Palacio, 2020).

Se han tomado en cuenta los requerimientos en en base a la especificación de requisitos de las partes interesadas o stakholders como permite la norma IEEE 29148 y se presentan a continuación: en la tabla 2, la lista de los interesados, en la tabla 3, los requerimientos de negocio y en la tabla 4 los requerimientos iniciales del sistema. Con el fin de seleccionar las más óptimas para este trabajo, ya

que según los estudios realizados no hay una única herramienta que aborde todos los requerimientos del cliente y por ello se recomienda realizar una combinación para obtener mejores resultados en los casos que fueran necesarios, posterior a eso realizar su instalación y las respectivas pruebas.

Tabla 2 *Se presenta la lista de interesados o Stackholders*

Lista de Stackholders
Operadores de redes móviles
Universidad Técnica del Norte
Msc. Diana Martínez
Ing. Samantha Mesa
Instituciones públicas y privadas

Nota: *Realizado por el autor*

Tabla 3 *Requerimientos de negocio, norma IEEE 29148*

StSR						
REQUERIMIENTOS DE NEGOCIO						
#	**REQUERIMIENTO**	**PRIORIDAD**			**RELACIÓN**	**VERIFI-CACIÓN**
		Alta	**Media**	**Baja**		
StRS 1	Herramientas de ingestión de software libre sin costo	☒	☐	☐		
REQUERIMIENTOS OPERACIONALES						
StRS 2	La ingestión de datos debe ser fiable y garantizada	☒	☐	☐		
StRS 3	Las herramientas deben adaptarse al ambiente de prueba propuesto	☒	☐	☐		
REQUERIMIENTOS DE USUARIOS						
StRS 4	Que manejen el tipo de archivos generados por la red móvil 5G	☒	☐	☐		

Nota: *Realizado por el autor*

Tabla 4 *Requerimientos iniciales del sistema, norma IEEE 29148*

Nota: *Realizado por el autor*

SySR						
REQUERIMIENTO DE FUNCIONES						
#	**REQUERIMIENTO**	**PRIORIDAD**			**RELACIÓN**	**VERIFI-CACIÓN**
		Alta	**Media**	**Baja**		
SySR 1	Las herramientas deben tener un buen rendimiento para la ingesta masiva de datos	☒	☐	☐	**StSR 2**	*S*
REQUERIMIENTO DE USO						
SySR 2	Estabilidad	☒	☐	☐		
SySR 3	Disponibilidad	☒	☐	☐		
SySR 4	Compatibilidad	☒	☐	☐		
SySR 5	Seguridad	☒	☐	☐		
REQUERIMIENTO DE PERFORMANCE						
SySR 6	Velocidad	☒	☐	☐		
SySR 7	Archivos procesados por segundo	☒	☐	☐		
SySR 8	Escalabilidad	☒	☐	☐		
SySR 9	Durabilidad del mensaje	☒	☐	☐		
REQUERIMIENTO DE INTERFACES						
SySR 10	Interfaz para visualizar el comportamiento de la herramienta de ingestión	☒	☐	☐		
REQUERIMIENTO DE MODOS/ESTADOS						
SySR 11	Se necesita que las herramientas se encuentren en modo de prueba	☒	☐	☐		
REQUERIMIENTO FÍSICOS						
SySR 12	Computador con 16 GB de memoria RAM y con un CPU Intel Core i7 con cuatro núcleos de 2.70 GHz	☒	☐	☐		

Es necesario conocer qué herramienta se adecúa a las necesidades de cada caso, teniendo en cuenta una solución open source y así poder evaluarlas. (Salas, 2017)

La elección de la tecnología a utilizar depende del tipo de procesamiento que se llevará a cabo, ya sea un procesamiento por lotes o tiempo real, además de la fuente de los datos. Existen varias herramientas que se han desarrollado para realizar esta tarea. (Sánchez, 2019)

De acuerdo a la información recopilada de diferentes fuentes como se menciona en (Matacuta & Popa, 2018) y (Sánchez, 2019) algunos artículos, libros y foros realizados acerca de las herramientas de ingestión de datos más utilizadas en Big Data, analizan las características y el rendimiento de las mismas, por lo que recomiendan: Flume, Kafka y Nifi ya que aseguran consistentemente buenos resultados, estas soportan archivos con formato XML y también se adaptan a las necesidades específicas basándose en indicadores de funcionalidad y rendimiento (Matacuta & Popa, 2018).

En otros estudios se determina que las herramientas más utilizadas para la recopilación de datos son Flume y Kafka (Cacho, 2014).

A continuación, en la tabla 5 y 6 se realiza la elección de herramientas en base a los requerimientos obtenidos a través de la norma IEEE 29148

Tabla 5 *Elección de herramientas en base a requerimientos de negocio de la norma IEEE 29148*

HERRAMIENTA	REQUERIMIENTOS STRS				VALORACION TOTAL
	1	2	3	4	
Flume	●	●	●	●	4
Nifi	●	●	○	●	3
Kafka	●	●	●	●	4
Hadoop	●	●	●	●	4
Sqoop	●	●	○	○	2
Spark	●	●	○	○	2
RabbitMQ	●	●	○	●	3
Amazon kinesis	○	●	○	●	2
Event Hubs	○	●	○	●	2
Cumple	●				
No cumple	○				

Nota: *Realizado por el autor*

Tabla 6 *Tabla 5 Elección de herramientas en base a requerimientos del sistema de la norma IEEE 29148*

HERRAMIENTA	REQUERIMIENTOS SySR										VALORACION TOTAL
	1	2	3	4	5	6	7	8	9	10	
Flume	●	●	●	●	●	●	●	●	●	●	10
Nifi	●	●	○	●	●	●	●	○	●	●	8
Kafka	●	●	●	●	●	●	●	●	●	●	10
Hadoop	●	●	●	●	●	●	●	●	●	●	10
Sqoop	●	○	●	●	●	●	●	○	●	○	7
Spark	●	●	●	●	●	○	●	●	●	○	8
RabbitMQ	●	●	●	○	●	●	●	●	●	●	9
Amazon kinesis	●	●	●	○	●	○	●	○	●	●	7
Event Hubs	●	●	●	○	●	●	●	○	●	●	8
Cumple	●										
No cumple	○										

Nota: *Realizado por el autor*

De acuerdo al análisis comparativo que se puede obtener de las tablas 5 y 6 de las herramientas de ingestión de datos, se determina que las tres herramientas que cumplen con todos los requerimientos que se plantean en este trabajo son: FLUME, KAFKA y HADOOP.

A continuación, se realiza una descripción de cada una de las herramientas de ingestión seleccionadas, para explicar su funcionamiento.

Kafka

Esta es la primera herramienta con la que se realizaron las pruebas, el clúster de Kafka requiere una instalación y configuración inicial y también otra configuración para los conectores para su correcto funcionamiento como se puede ver en el anexo B.

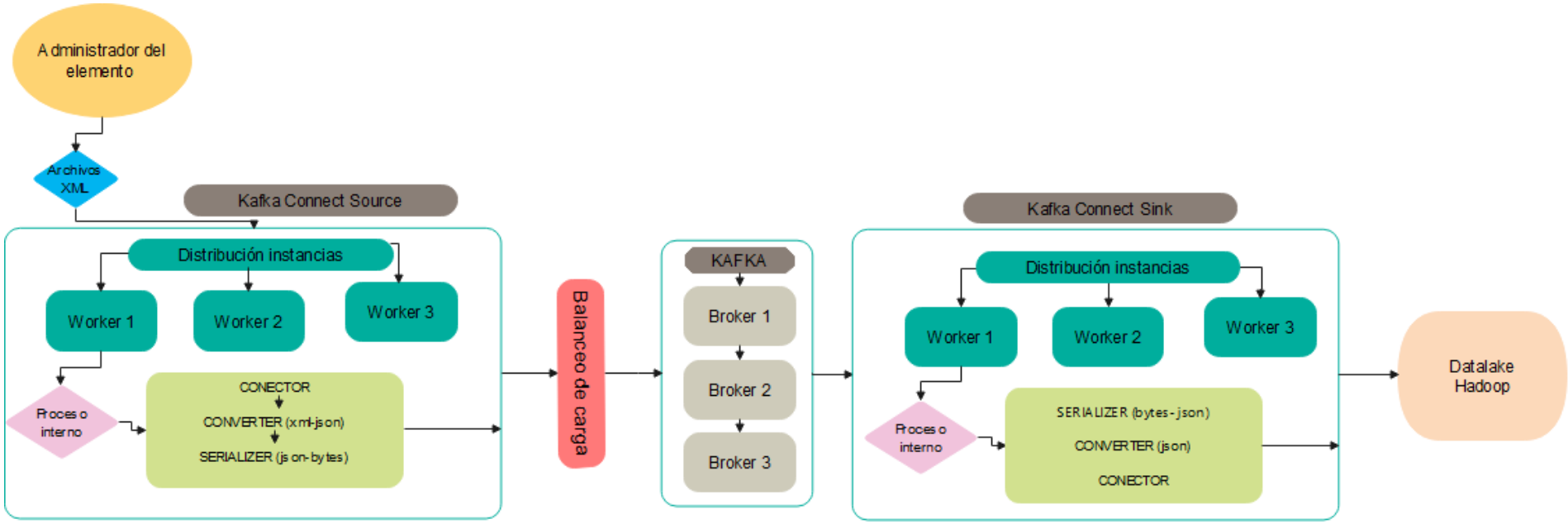
En la figura 5 se muestra el proceso que siguen los archivos xml desde que salen del administrador del elemento hasta llegar al datalake utilizando la herramienta Kafka.

En primer lugar, se tiene el elemento de red donde se originan los archivos xml, al cual se conecta el conector de Kafka o llamado también connect source cumple la función de un productor ya que lee un archivo de un origen de datos y produce mensajes a Kafka.

Es importante indicar que los archivos que ingresan al Kafka connect source, cumplen con un proceso interno en los workers los cuales de manera interna realizan otros procesos de conversión de xml a formato json, y luego el proceso de serialización que cambia el formato de json a bytes, se agregan cabeceras y se realiza un balanceo de carga para posterior a eso enviar el mensaje al clúster de Kafka.

En el clúster de Kafka se tienen los brokers, los cuales reciben los mensajes de los workers y los envían al siguiente conector llamado connect sink, el cual realiza el mismo proceso que el connect source, pero de manera inversa, en sus workers se reciben los mensajes en bytes, y se cambia su formato a json, este conector cumple la función de consumidor ya que lee los mensajes de Kafka y los envía a un repositorio destino, en este caso el datalake en Hadoop.

Figura 5 *Proceso de ingestión de datos utilizando la herramienta Kafka*



Nota: Realizado por el Autor

Flume

Esta es la segunda herramienta con la que se han realizado las pruebas, Flume recopila, agrega y mueve datos desde diversas fuentes hasta almacenamientos de datos centralizados, para ello es importante comprender cómo se comporta y qué criterios utiliza, en Flume se utilizan varios conceptos para comprender su funcionamiento, un evento Flume se define como una unidad de flujo de datos que tiene una carga útil de bytes y un conjunto opcional de atributos de cadena. Un agente de Flume es un proceso que aloja los componentes a través de los cuales los eventos fluyen desde una fuente externa al siguiente destino (salto). La configuración de Flume se almacena en un archivo de configuración local. Este es un archivo de texto que sigue el formato de archivo de propiedades de Java (The Apache Software Foundation, n.d.).

Una fuente de Flume consume los eventos que le envía una fuente externa. La fuente externa envía eventos a Flume en un formato que es reconocido por la fuente de Flume de destino. Cuando una fuente de Flume recibe un evento, lo almacena en uno o más canales. El canal es una tienda pasiva que mantiene el evento hasta que es consumido por un sumidero Flume. El canal de archivos es un ejemplo: está respaldado por el sistema de archivos local. El receptor elimina el evento del canal y lo coloca en un repositorio externo como HDFS (a través del receptor Flume HDFS) o lo reenvía a la fuente Flume del siguiente agente Flume (siguiente salto) en el flujo. La fuente y el receptor dentro del agente dado se ejecutan de forma asincrónica con los eventos organizados en el canal (The Apache Software Foundation, n.d.).

- Evento: Son las unidades de datos transportadas por el agente de Flume.
- Agente: Contenedor para alojar subcomponentes que permiten mover los eventos.
- Source: Receptor de eventos.
- Canal: Buffer de eventos

La arquitectura de Flume tiene tres niveles:

Nivel de agente: este es el nivel donde se encuentran los agentes de Flume junto con las fuentes que contienen datos que deben moverse.

Nivel de recopilador: este es el nivel en el que se recopilan los datos del nivel de agente mediante varios recopiladores y luego se reenvían a la siguiente capa.

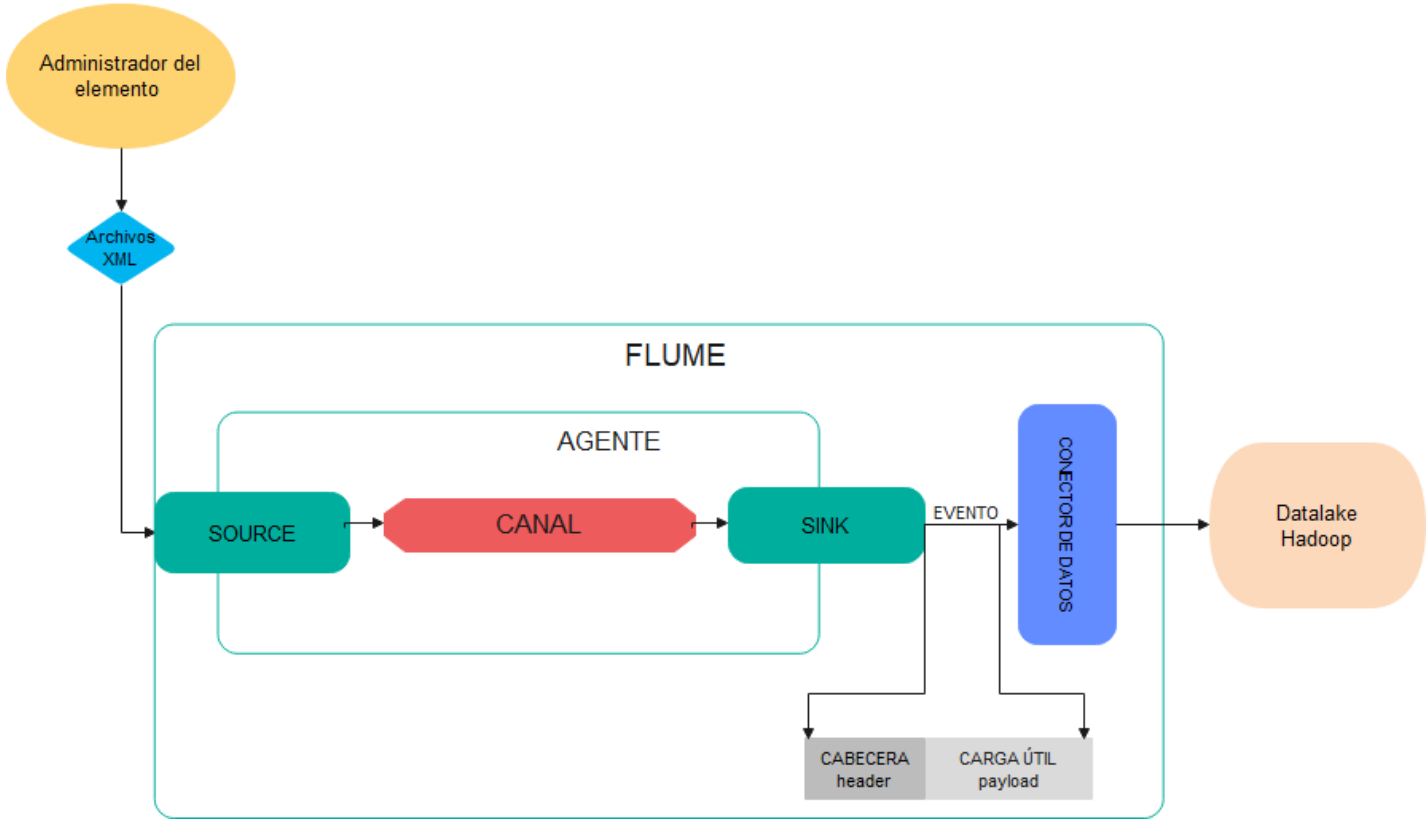
Nivel de almacenamiento: este es el nivel donde los datos del nivel del recopilador fluyen finalmente y se almacenan. Esto tendrá sistemas de archivos como HDFS donde se almacenan los datos, su configuración se muestra en el anexo C.

En la figura 6 se muestra el proceso que siguen los archivos xml desde que salen del administrador del elemento hasta llegar al datalake utilizando la herramienta Flume.

Se tiene el elemento de red donde se originan los archivos xml, el cual se conecta al agente de Flume, dentro del agente se realiza un proceso interno para poder mover los eventos.

El componente source recibe los archivos xml, los convierte en eventos (array de bytes), que son las unidades de datos que transporta el agente y los almacena en el canal hasta que sean consumidos por el sink, este toma eventos del canal y los transmite hacia el siguiente componente, se agregan cabeceras a la carga útil, se transforman nuevamente a su formato de origen y se envían al datalake.

Figura 6 *Proceso de ingestión de datos utilizando la herramienta Flume*

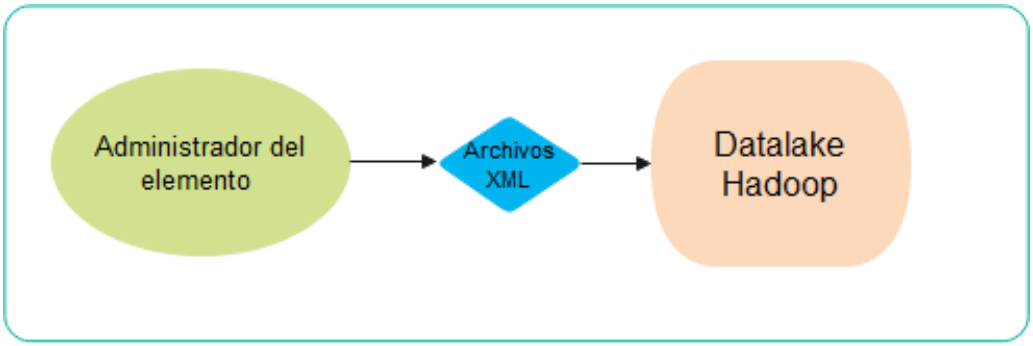


Nota: Realizado por el Autor

Hadoop

Para la tercera herramienta se ha utilizado Hadoop, en esta herramienta no se requiere una configuración compleja como en las anteriores, los archivos xml se envían directamente desde el elemento de red (fuente) hacia el datalake.

Figura 7 *Proceso de ingestión de datos utilizando la herramienta Hadoop*



Nota: Realizado por el Autor

CAPÍTULO IV: RESULTADOS Y ANÁLISIS

En este capítulo se muestra un detalle de los resultados obtenidos en las pruebas realizadas con cada herramienta de ingestión de datos y su respectivo análisis.

Es importante mencionar que para realizar las pruebas se han utilizado 16 GB en memoria RAM, con un CPU Intel Core i7 con cuatro núcleos de 2.70 GHz, un elemento de red simulado en donde se originan todos los archivos XML con distintos tamaños y el repositorio Hadoop (Datalake) para el almacenamiento de los archivos.

Para la instalación de las herramientas se utilizaron las siguientes versiones: para Kafka la versión 2.6.0, para Flume la versión 1.9.0 y para Hadoop la versión 3.2.2.

Se utilizó un archivo 5G de una red de prueba, el mismo que ha sido multiplicado hasta obtener 20 muestras, cada una contiene una cantidad de archivos que van de manera ascendente, desde un archivo

hasta 8000 archivos, lo que permitió ver el comportamiento de las herramientas y a su vez determinar parámetros importantes como:

- Tiempo de ingesta, es la hora en que se envía cada muestra desde el elemento de red hacia el datalake en hadoop.
- Tiempo en datalake, es la hora en que termina el proceso de ingesta y los archivos ya se encuentran en el datalake de hadoop.
- Tiempo, medido en segundos, en este campo se resta el tiempo de ingesta del tiempo en datalake para determinar cuánto se tardó cada muestra.
- Tamaño de muestra, se midió el tamaño de cada muestra una vez sé que encuentran en el datalake.
- Consumo CPU, se determinó cuánto porcentaje de CPU fue requerido con cada muestra, se tuvieron variaciones de acuerdo a su tamaño.
- Consumo RAM, se determinó cuánta memoria en gigabytes fue requerida con cada muestra, se tuvieron variaciones de acuerdo a su tamaño.

Resultados de la herramienta Kafka

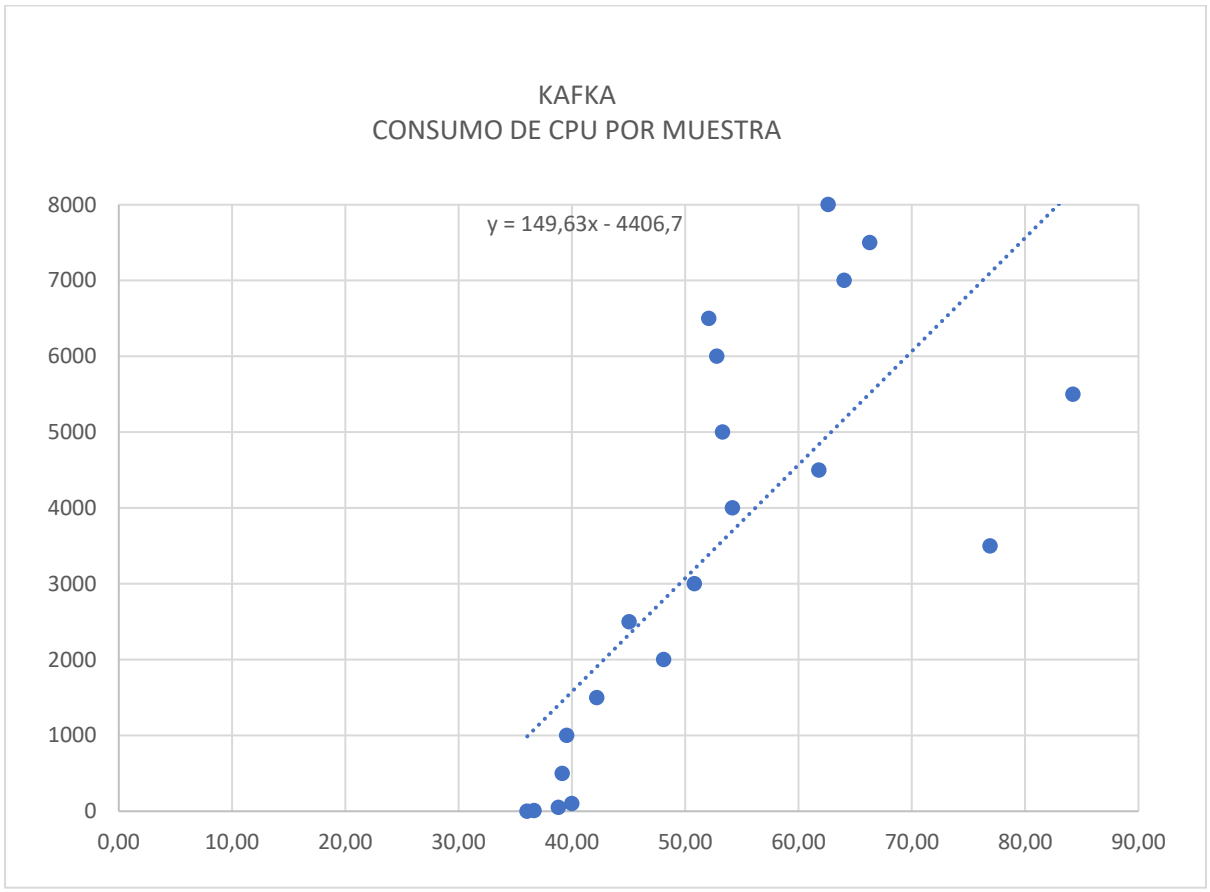
A continuación, en la tabla 4 se muestran los resultados obtenidos tanto en tiempo de ingestión como consumo de CPU y memoria para diferentes tamaños de archivos utilizando la herramienta Kafka.

Tabla 7 *Resultados obtenidos con la herramienta Kafka*.

HERRAMIENTA	MUESTRA	ARCHIVOS (xml)	Tiempo de ingesta	Tiempo en datalake	TIEMPO (s)	TAMAÑO	CONSUMO CPU %	CONSUMO RAM (15.5 GB)
KAFKA	1	1	12:31:58	12:23:07	9	259.7 KB	36.05	10.4
	2	10	12:43:14	12:43:22	8	2.6 MB	36.67	11.1
	3	50	12:50:45	12:51:13	28	13.0 MB	38.8	11.4
	4	100	12:57:51	12:58:41	50	26.0 MB	40.0	11.5
	5	500	13:07:14	13:11:24	250	129.8 MB	39.15	11.0
	6	1000	13:40:08	13:48:28	500	260.0 MB	39.55	11.3
	7	1500	13:56:38	14:09:05	747	389.9 MB	42.2	11.6
	8	2000	14:24:14	14:41:00	1006	519.9 MB	48.1	12.5
	9	2500	15:27:12	15:48:01	1249	649.9 MB	45.05	12.3
	10	3000	16:02:20	16:27:14	1494	779.9 MB	50.8	12.4
	11	3500	16:32:43	17:01:47	1744	909.9 MB	76.9	12.2
	12	4000	17:14:49	17:48:08	1999	1000 MB	54.17	11.8
	13	4500	17:56:51	18:34:21	2250	1.2 GB	61.8	11.9
	14	5000	18:42:33	19:24:23	2510	1.3 GB	53.3	11.9
	15	5500	19:37:29	20:23:31	2762	1.4 GB	84.22	11.7
	16	6000	20:32:31	21:22:45	3014	1.6 GB	52.8	12.0
	17	6500	21:42:27	22:36:50	3263	1.7 GB	52.07	12.3
	18	7000	22:58:36	23:57:11	3515	1.8 GB	64.02	12.1
	19	7500	0:14:01	1:16:48	3767	1.9 GB	66.3	12.3
	20	8000	15:52:32	16:59:25	4013	2.1 GB	62.62	10.1

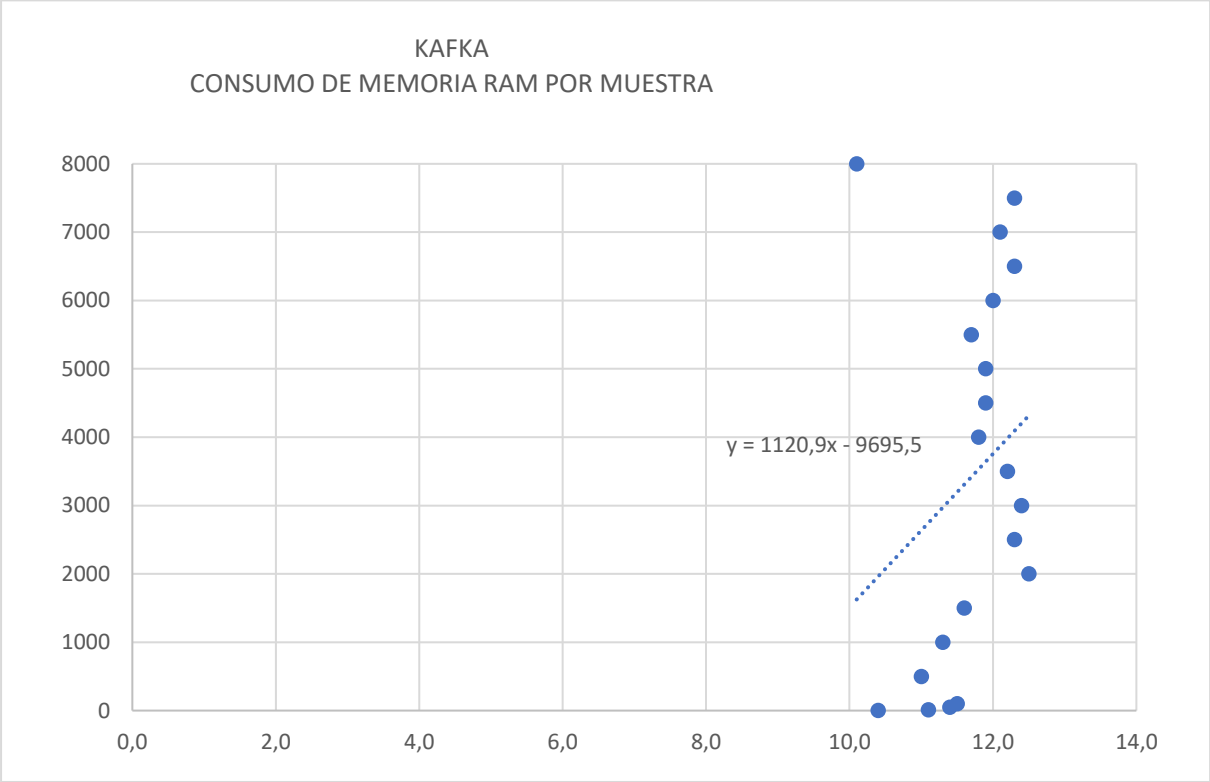
A continuación, se muestran las gráficas que representan el consumo de CPU en la figura 8 y de memoria RAM en la figura 9 de la herramienta Kafka, en donde se pueden observar que a medida que aumenta el número de archivos también aumenta el consumo tanto de CPU como de memoria RAM, al representar los datos en las gráficas, se puede ver que tienen una tendencia lineal, esto quiere decir que existe una relación directa entre el tamaño de las muestras con el consumo de CPU y de memoria RAM y que existe una fuerte dependencia lineal entre ellos.

Figura 8 *Consumo de CPU de la herramienta Kafka.*



Nota: Realizado por el Autor

Figura 9 *Consumo de memoria RAM de la herramienta Kafka.*



Nota: Realizado por el Autor

Resultados de la herramienta Flume

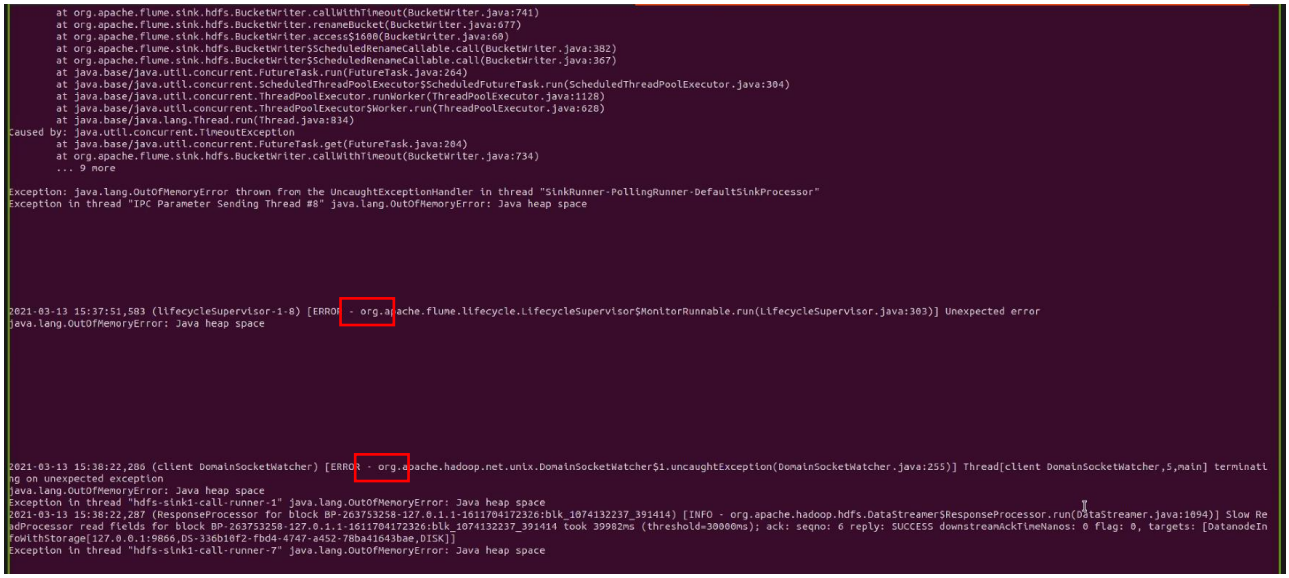
A continuación, en la tabla 5 se muestran los resultados obtenidos tanto en tiempo de ingestión como consumo de CPU y memoria para diferentes tamaños de archivos utilizando la herramienta Flume.

Tabla 8 *Resultados obtenidos con la herramienta Flume.*

HERRAMIENTA	MUESTRA	ARCHIVOS (xml)	Tiempo de ingesta	Tiempo en datalake	TIEMPO (s)	TAMAÑO	CONSUMO CPU %	CONSUMO RAM (15.5 GB)	ARCHIVOS	TAMAÑO
FLUME	1	1	19:29:42	19:29:43	1	285 KB	46.05	7.73 GB	1	285 KB
	2	10	19:36:01	19:36:07	6	2.78 MB	43.85	7.81 GB	10	2.78 MB
	3	50	19:41:58	19:42:17	19	13.91 MB	71.00	8.29 GB	50	13.91 MB
	4	100	19:44:50	19:46:22	92	27.83 MB	66.92	9.02 GB	100	27.83 MB
	5	500	19:50:18	19:57:04	406	139.16 MB	84.82	11.1 GB	543	151.13 MB
	6	1000	20:01:56	20:16:00	844	278.33 MB	89.27	11.3 GB	1102	306.72 MB
	7	1500	20:59:25	21:18:08	1123	417.50 MB	68.77	12.5 GB	1592	443.11 MB
	8	2000	21:27:45	21:51:58	1453	556.67 MB	63.85	12.5 GB	2122	590.63 MB
	9	2500	19:52:10	20:22:07	1797	695.84 MB	62.67	11.3 GB	2657	739.54 MB
	10	3000	20:26:26	20:59:48	2002	835.01 MB	64.25	12.2 GB	3193	888.73 MB
	11	3500	21:05:04	21:45:58	2454	974.18 MB	58.22	13.1 GB	3744	1.01 GB
	12	4000	11:30:43	12:28:17	3454	1.08 GB	64.72	13.2 GB	4357	1.18 GB
	13	4500	12:31:29	13:34:14	3765	1.22 GB	64.72	13.3 GB	4932	1.34 GB
	14	5000								
	15	5500								
	16	6000								
	17	6500								
	18	7000								
	19	7500								
	20	8000								

No hay datos a partir de la muestra 14 porque el procesamiento de la memoria RAM no fue suficiente, la herramienta necesita más recursos para trabajar correctamente, las pruebas se realizaron con éxito hasta la muestra 13 y de ahí en adelante se generaron errores como se muestra en la figura 10.

Figura 10 *Errores en las pruebas con la herramienta Flume*

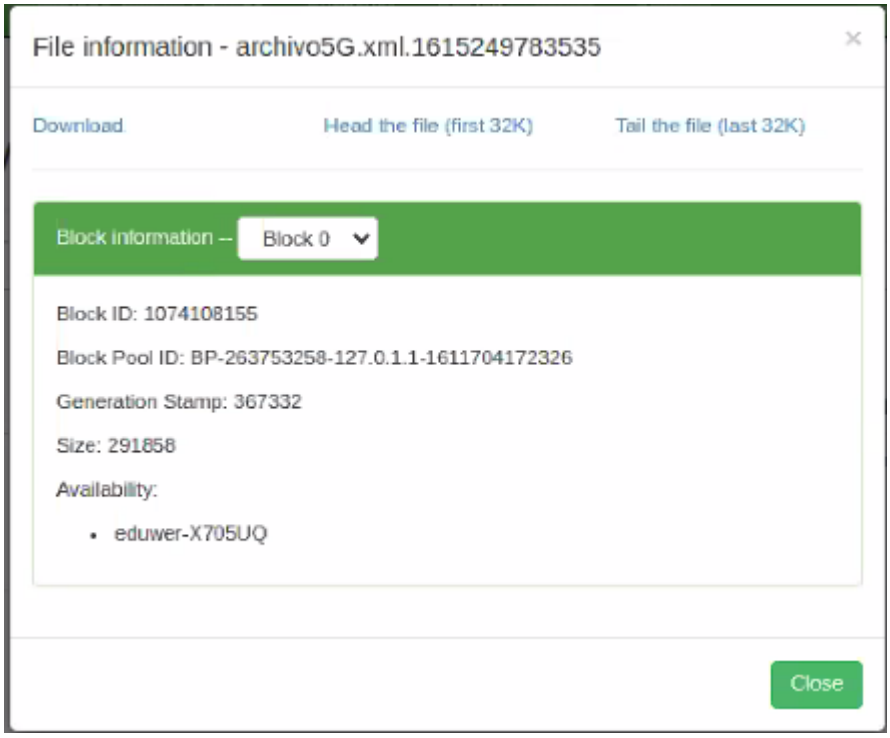


Nota: Realizado por el Autor

El tamaño original de los archivos XML de prueba es 191 KB y de acuerdo a las pruebas realizadas, los archivos que fueron ingestados por la herramienta Flume, se ingestan con su tamaño original, pero como se muestra en la figura 11 al repositorio en Hadoop llegan con un mayor tamaño de 291858 bytes es decir 285 KB, esto debido a que los conectores de Flume aumentan cabeceras en el agente.

De acuerdo al estado del arte en Flume pueden duplicarse archivos, pero no perderse.

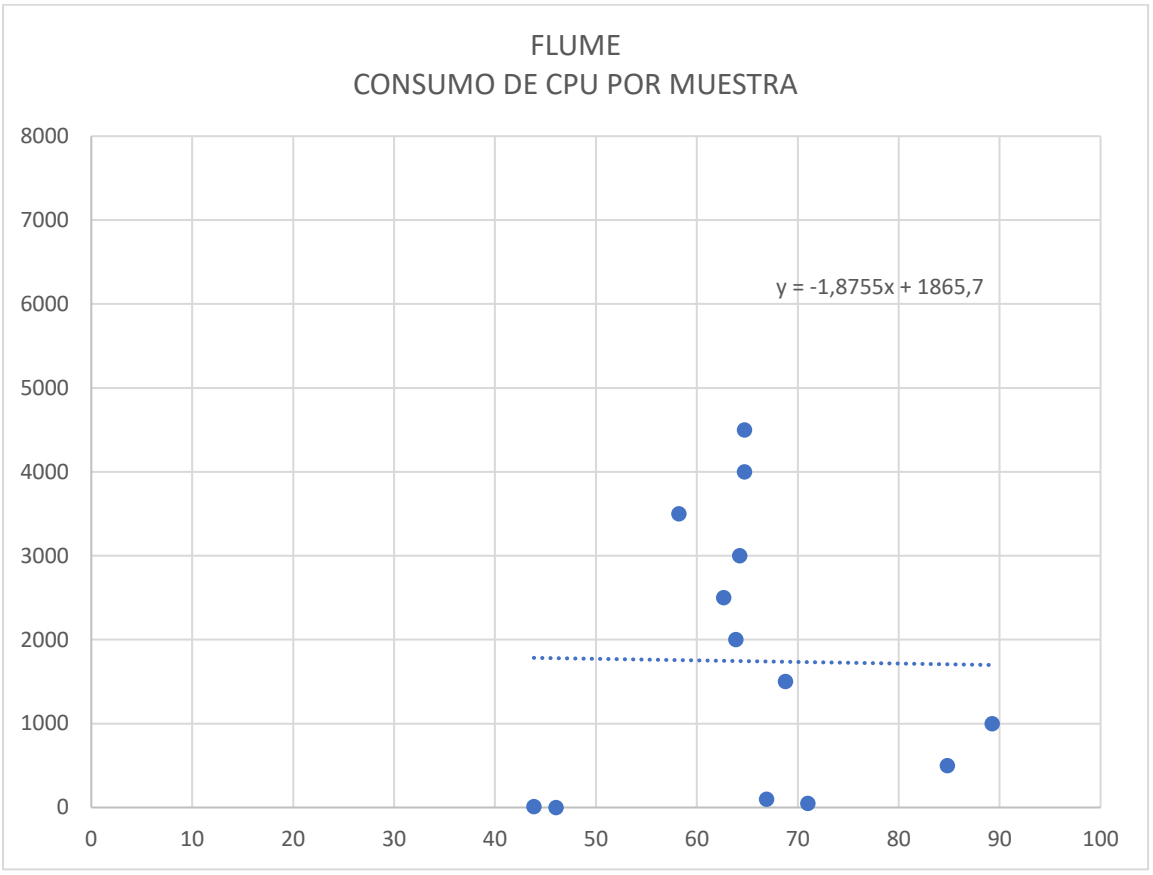
Figura 11 *Archivo xml ingestado por la herramienta Flume descargado desde el datalake.*



Nota: Realizado por el Autor

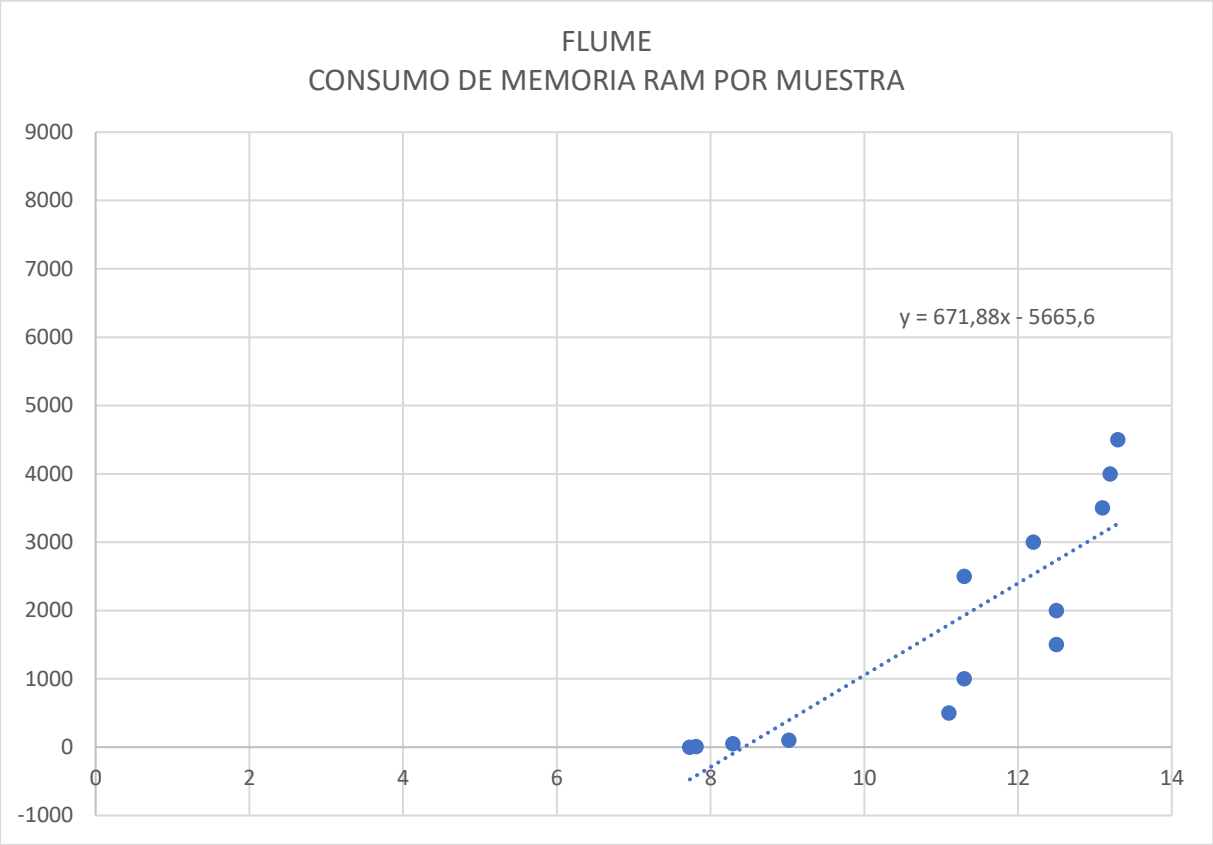
A continuación, se muestran las gráficas que representan el consumo de CPU en la figura 12 y de memoria RAM en la figura 13 de la herramienta Flume, en donde se pueden observar que a medida que aumenta el número de archivos también aumenta el consumo tanto de CPU como de memoria RAM, al representar los datos en las gráficas, se puede ver que tienen una tendencia lineal, esto quiere decir que existe una relación directa entre el tamaño de las muestras con el consumo de CPU y de memoria RAM y que existe una fuerte dependencia lineal entre ellos.

Figura 12 *Consumo de CPU de la herramienta Flume.*



Nota: Realizado por el Autor

Figura 13 *Consumo de memoria RAM de la herramienta Flume.*



Nota: Realizado por el Autor

Resultados de la herramienta Hadoop

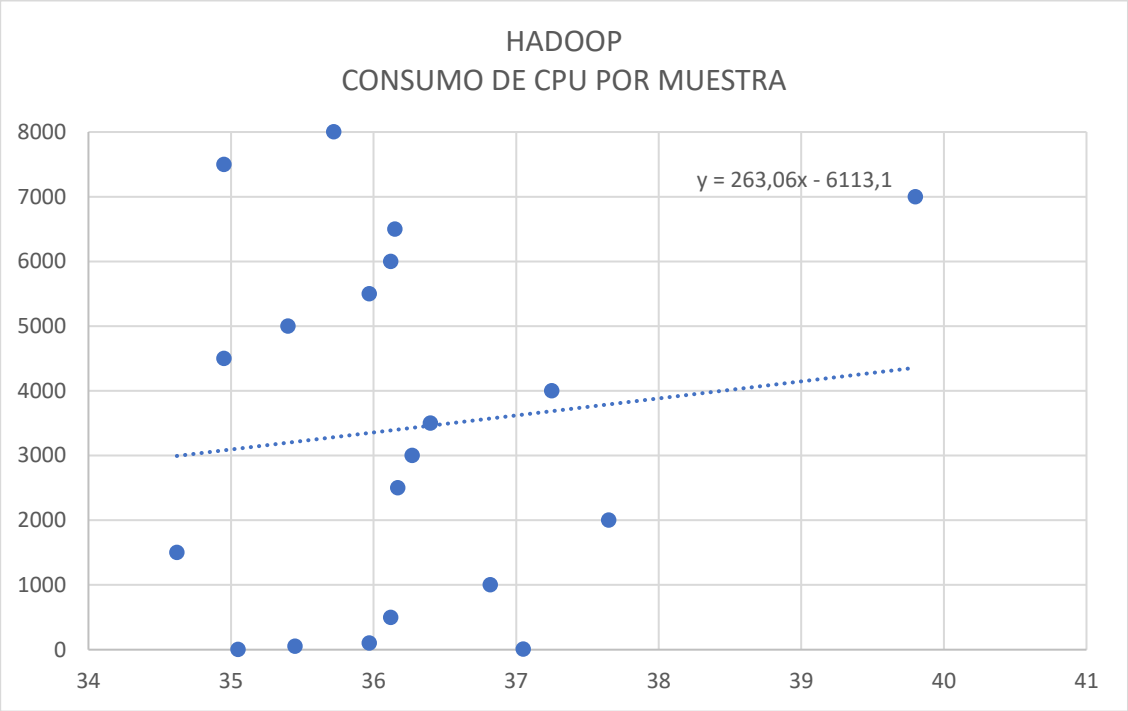
A continuación, en la tabla 6 se muestran los resultados obtenidos tanto en tiempo de ingestión como consumo de CPU y memoria para diferentes tamaños de archivos utilizando la herramienta Hadoop.

Tabla 9 *Resultados obtenidos con la herramienta Hadoop.*

HERRAMIENTA	MUESTRA	ARCHIVOS (xml)	Tiempo de ingesta	Tiempo en datalake	TIEMPO (s)	TAMAÑO	CONSUMO CPU %	CONSUMO RAM (15.5 GB)
HADOOP	1	1	21:39:20	21:39:22	2	191 KB	35.05	4.84
	2	10	21:44:10	21:44:14	4	1910 KB	37.05	4.85
	3	50	21:49:24	21:49:28	4	9550	35.45	4.85
	4	100	21:50:42	21:50:51	9	19100	35.97	4.85
	5	500	21:57:15	21:57:52	37	95500	36.12	4.88
	6	1000	21:58:57	21:59:59	62	191000	36.82	4.95
	7	1500	22:02:03	22:03:32	89	286500	34.62	4.98
	8	2000	22:06:38	22:09:05	147	382000	37.65	5.04
	9	2500	22:10:50	22:13:09	139	477500	36.17	5.02
	10	3000	22:14:50	22:17:50	180	573000	36.27	5.00
	11	3500	22:18:59	22:22:36	217	668500	36.4	5.02
	12	4000	22:23:51	22:27:53	242	764000	37.25	5.10
	13	4500	22:28:53	22:34:15	322	859500	34.95	5.09
	14	5000	22:35:29	22:40:13	284	955000	35.4	5.07
	15	5500	22:41:31	22:47:25	354	1050500	35.97	5.16
	16	6000	22:49:01	22:56:04	423	1146000	36.12	5.12
	17	6500	22:57:20	23:05:16	476	1241500	36.15	5.11
	18	7000	23:06:03	23:14:28	505	1337000	39.8	5.27
	19	7500	23:15:46	23:23:01	435	1432500	34.95	5.28
	20	8000	23:24:35	23:33:06	511	1528000	35.72	5.26

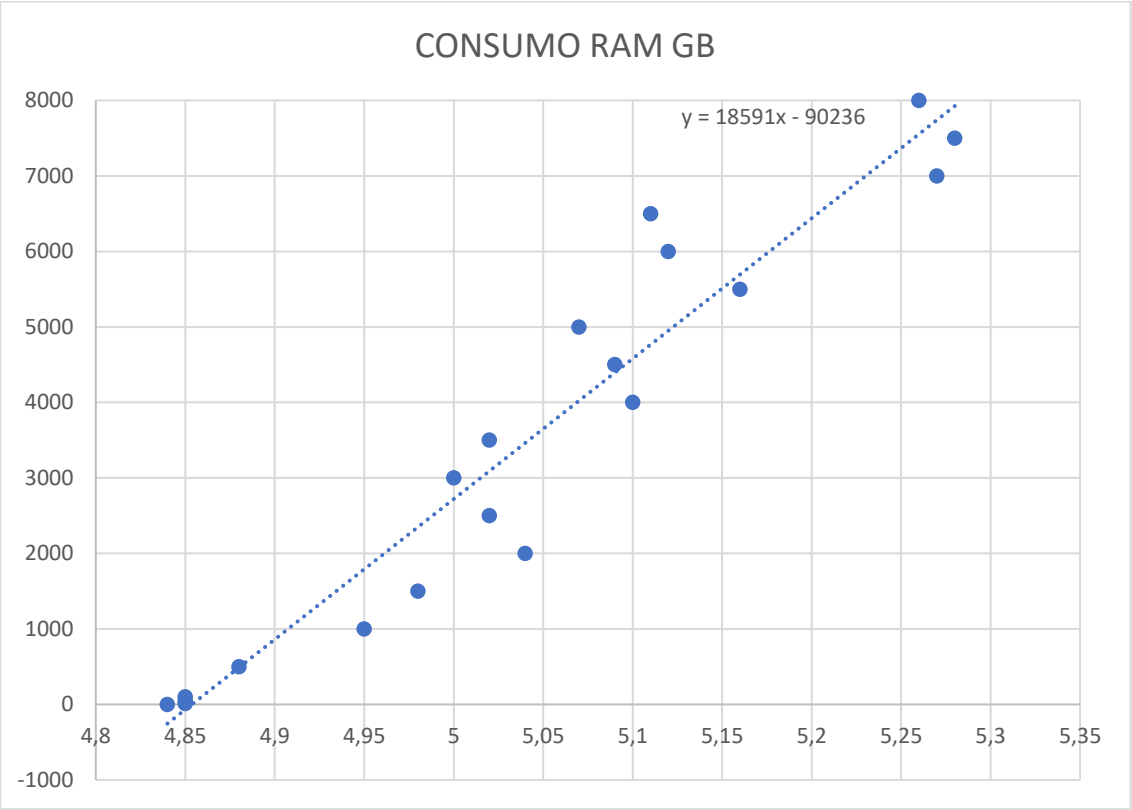
A continuación, se muestran las gráficas que representan el consumo de CPU en la figura 14 y de memoria RAM en la figura 15 de la herramienta Flume, en donde se pueden observar que a medida que aumenta el número de archivos también aumenta el consumo tanto de CPU como de memoria RAM, al representar los datos en las gráficas, se puede ver que tienen una tendencia lineal, esto quiere decir que existe una relación directa entre el tamaño de las muestras con el consumo de CPU y de memoria RAM y que existe una fuerte dependencia lineal entre ellos.

Figura 14 *Consumo de CPU de la herramienta Hadoop.*



Nota: Realizado por el Autor

Figura 15 *Consumo de memoria RAM de la herramienta Hadoop.*



Nota: Realizado por el Autor

Análisis de resultados

Las tres herramientas permitieron ingestar archivos XML que son los que se manejan en una red móvil 5G, Kafka, Flume y Hadoop tienen persistencia, confiabilidad, recogen, agregan y mueven grandes volúmenes de datos, esto permite que todos los datos de la red móvil 5G sean ingestados de manera eficiente y a la vez la red móvil sea monitoreada constantemente, a continuación, se presentan tres tablas comparativas sobre las ventajas y desventajas de Kafka, Flume y Hadoop respectivamente para trabajar con la red móvil 5G.

Tabla 10 *Ventajas y desventajas de la herramienta Kafka para la red móvil 5G.*

VENTAJAS	DESVENTAJAS
Tuvo un correcto funcionamiento, alto rendimiento y baja latencia en todas las pruebas realizadas con los archivos 5G sin generar errores, esto es de gran ayuda para saber el estado de la red móvil 5G en tiempo real.	Se requiere la instalación de Kafka y de los conectores por separado, también requiere la instalación de Java y ZooKeeper para funcionar.
Requiere menos procesamiento de memoria RAM, que la herramienta Flume para cumplir con la ingestión de datos 5G.	Los archivos son cambiados de formato al pasar por el conector consumidor, alterando también su tamaño, se requiere otro proceso adicional para volver a su formato original que demandaría tiempo.
El nodo intermediario (Kafka) también proporciona almacenamiento, retiene temporalmente los datos y tiene redundancia, en el caso que hubiera fallos.	En un ambiente real se requiere más capacidad de procesamiento del que se utilizó en las pruebas para probar su funcionamiento a gran escala.

Tabla 11 *Ventajas y desventajas de la herramienta Flume para la red móvil 5G.*

VENTAJAS	DESVENTAJAS
Tuvo un correcto funcionamiento, alto rendimiento y baja latencia en las pruebas que pudieron ser realizadas, es decir hasta la muestra 13 con los archivos 5G sin generar errores, en tiempo real es muy útil para el monitoreo de la red móvil 5G.	Requiere mayor memoria RAM, se tuvieron dificultades al aumentar el tamaño de las muestras, generó errores.
	Está diseñada para que los eventos vayan a un destino específico, aunque se pueden configurar varios agentes, es decir tiene dificultad para escalar.

Requiere una sola instalación y configuración, tanto para Flume como de los conectores. Flume tardó menos tiempo que Kafka en la ingestión de los archivos 5G. En su archivo de configuración, permite definir el flujo de ejecución de manera que se puede indicar los componentes por los que va pasando la ejecución.	Cuando se requiere que la latencia sea baja, la cantidad de datos que puede manejar es limitada.

Tabla 12 *Ventajas y desventajas de la herramienta Hadoop para la red móvil 5G.*

VENTAJAS	ESVENTAJAS
La configuración para que realice la ingestión desde el servidor que contiene archivos de la red móvil 5G es sencilla ya que en este trabajo de investigación es simulado. Tuvo un correcto funcionamiento, alto rendimiento y baja latencia en todas las pruebas realizadas con los archivos 5G sin generar errores. No requiere el uso de conectores ya que los archivos son ingestados directamente en el datalake. Almacena grandes cantidades de datos, estos no tienen alteraciones en su formato ni en su tamaño	No es lo mismo un ambiente de prueba que un ambiente real, esto hace que en su configuración se tengan que considerar otros aspectos importantes como la seguridad. Tiene actualizaciones constantes.

Permite hacer consultas con tiempos bajos de
respuesta.

En base a las pruebas realizadas en el proceso de ingestión de los archivos de una red móvil 5G, se pudo evaluar el desempeño de cada herramienta seleccionada y esto a su vez permite que se puedan analizar algunos aspectos:

- Se determinó el tiempo de ingestión, consumo de CPU y consumo de memoria RAM de los archivos 5G en base al tamaño de las diferentes muestras en donde: la herramienta Flume tarda menos que Kafka en casi todas las pruebas que pudieron ser realizadas, sin embargo las pruebas con Flume no pudieron ser concluidas, ya que esta herramienta de acuerdo a los resultados obtenidos consume más recursos tanto de CPU como de memoria RAM que Kafka y Hadoop, y por ende no se pudieron realizar las pruebas con todas las muestras ya que a partir de la muestra 14 que es la que contiene 5000 archivos, se generaron errores.

- Con las herramientas Kafka y Hadoop se pudieron realizar todas las pruebas de manera eficiente y sin errores, Kafka tarda un poco más que las otras dos herramientas, así también la herramienta Hadoop es la que menos consumo de CPU, memoria RAM y tiempo de ingestión tiene ya que no tiene muchos elementos de configuración y la ingestión se ha realizado directamente desde el servidor que contiene los archivos 5G.

- En base a los resultados obtenidos en esta investigación, considerando que las pruebas con esta herramienta se realizaron con éxito y que en un ambiente real se requiere una tecnología que cuente con un sistema robusto, distribuido y tolerante a fallos se recomienda la herramienta Kafka como la herramienta más eficiente para el proceso de ingestión de archivos 5G en la capa ingestión.

CAPÍTULO V: CONCLUSIONES Y RECOMENDACIONES

Conclusiones

- Se evaluó el desempeño de una arquitectura de Big Data para 5G a nivel de la capa de ingestión, usando tres aplicaciones de recolección de datos Kafka, Flume y Hadoop, las mismas que han trabajado de manera eficiente con archivos xml 5G que son que se han utilizado para realizar las pruebas en este trabajo de investigación.

- Para la selección de las herramientas de ingestión de datos se realizó un análisis en base indicadores de funcionalidad y de rendimiento lo que permitió tener un panorama claro de la selección realizada.

- Se realizaron pruebas de estrés del sistema, con esto se pudo evaluar el desempeño de cada herramienta seleccionada, se determinó el tiempo de ingestión de los archivos 5G en base al tamaño de las diferentes muestras en donde Flume tarda menos que Kafka en todas las pruebas que pudieron ser realizadas.

- De acuerdo a los resultados se evidencia que Flume requiere más procesamiento que Kafka y Hadoop, no se pudieron concluir las pruebas con todas las muestras, a partir de la muestra 14 que contiene 5000 archivos se generaron errores, con las herramientas Kafka y Hadoop se pudieron realizar todas las pruebas de manera eficiente y sin errores, considerando que en un ambiente real se requiere una tecnología que cuente con un sistema distribuido y tolerante a fallos se recomienda la herramienta Kafka como la herramienta más eficiente para el proceso de ingestión de archivos 5G en la capa ingestión.

Recomendaciones

- Tener en cuenta los requerimientos previos de memoria RAM y almacenamiento que cada herramienta necesita, ya que las pruebas con las tres herramientas se deben realizar en el mismo ambiente.

- Las herramientas utilizadas requieren la instalación previa de Java y también de otros servicios como ZooKeeper, se debe tener en cuenta esto antes de la configuración para que funcionen correctamente.

- Se recomienda verificar el correcto funcionamiento de cada herramienta previo a las pruebas ya que al procesar datos de diferentes tamaños se estresa al sistema y se van a ir generando errores.

REFERENCIAS BIBLIOGRAFICAS

3rd Generation Partnership Project. (2020). Extensible Markup Language (XML) file format definition. In *FIle markerPro 15* (Issue Xml, p. 4). https://fmhelp.filemaker.com/help/12/fmp/es/html/import_export.17.31.html

3rd Generation Partnership Project. (2021). *Technical Specification Group Services and System Aspects; System architecture for the 5G System (5GS); Stage 2 (Release 17)*. https://www.3gpp.org/ftp/Specs/archive/23_series/23.501/

Algorri Álvarez, M. (2017). *Caracterización de tecnologías de procesamiento de datos en streaming sobre una arquitectura orientada al dato*. 1–52.

ARCOTEL. (2020). Infraestructura Y Cobertura. *Boletin Estadistico*. https://www.arcotel.gob.ec/wp-content/uploads/2015/01/BoletinEstadistico-May2020-SMA-CoberturaInfraestructura.pdf

Barreno, D., Carrión, P., & Tenecora, I. (2016). *EVOLUCIÓN DE LA TECNOLOGÍA MOVIL CAMINO A 5G*.

Cacho, R. (2014). Tu cuenta tiene una tarjeta de crédito y no es necesario realizar ninguna acción. Se cargará el importe pendiente en la tarjeta archivada en los próximos 10 días. In *Zaguan.Unizar.Es*.

Careaga, J. (2017). *Arquitectura Lambda vs Arquitectura Kappa*. http://www.i2ds.org/wp-content/uploads/2017/01/i2ds.org_ArquitecturasKL.pdf

Díaz, J. (2016). La evolución de Internet y las tecnologías móviles. *Bit y Byte*, *4*, 19–20. http://redestelematicas.com/historia-de-internet-nacimiento-y-evolucion/

Díaz Zayas, A., Merino Gómez, P., & Rivas Tocado, F. J. (2017). *3GPP NB-IoT, tecnología y herramientas de medida*. *Jitel*, 310–317. https://doi.org/10.4995/jitel2017.2017.6577

Fernández, F. (2017). *Arquitectura Big Data de ingesta en Real Time*. 94.

González, J., & Salamanca, O. (2016). EL CAMINO HACIA LA TECNOLOGÍA 5G. *Universidad Privada Dr. Rafael Belloso Chacín, URBE, Venezuela*, 22.

GSMA. (2020). *La Economía móvil en América Latina 2020*. https://www.gsma.com/mobileeconomy/wp-

content/uploads/2020/12/GSMA_MobileEconomy2020_LATAM_Esp.pdf

Gutiérrez Álvaro, J. (2019). ¿a Las Puertas Del 5G? *Bit*, *211*, 57–60.

Hu, X., Liu, C., Liu, S., You, W., Li, Y., & Zhao, Y. (2019). A systematic analysis method for 5g non-access stratum signalling security. *IEEE Access*, *7*, 125424–125441. https://doi.org/10.1109/ACCESS.2019.2937997

Iñigo, N. (2020). Diseño de una plataforma Big Data para predicción de patologías a partir de resultados médicos. *Zaguan.Unizar.Es*, 70.

ITU. (2017). Abrir sendas. *22 Anual Spectrum Summit*, 36. https://www.itu.int/en/itunews/Documents/2017/2017-02/2017_ITUNews02-es.pdf

Jaramillo, N., Ochoa, A., Páez Alexander Peña, W. Y., Páez, W., & Alexander Peña, Y. (2017). *TECNOLOGÍA 5G 5G Technology I. DESARROLLO*. *4*, 41–45. https://doi.org/10.21017/rimci.2017.v4.n8.a31

Kamakhya Narain Singh, R. K. B. and J. K. M. (2018). *Big Data Ecosystem: Review on Architectural Evolution* (Vol. 2). https://doi.org/10.1201/9781420057331.ch7

Magalhães, R. (n.d.). *5G Y IOT TENDENCIAS E APLICACIONES*.

Maldonado, D. (2018). *4 Razones para automatizar la ingesta de datos*. http://www.icorp.com.mx/blog/automatizar-ingesta-de-datos/

Martinez, D., Navarrete, R., & Lujan, S. (2020). Development and Evaluation of a Big Data Framework for Performance Management in Mobile Networks. *IEEE Access*, *8*. https://doi.org/10.1109/ACCESS.2020.3045175

Matacuta, A., & Popa, C. (2018). Big Data Analytics: Analysis of Features and Performance of Big Data Ingestion Tools. *Informatica Economica*, *22*(2/2018), 25–34. https://doi.org/10.12948/issn14531305/22.2.2018.03

MENDOZA, G. A. T. (2016). *DISEÑO DE UN OPERADOR MÓVIL VIRTUAL CONVERGENTE ENTRE LAS REDES DE TERCERA GENERACIÓN DE LOS OPERADORES MÓVILES DE RED Y LA TECNOLOGIA DE EVOLUCIÓN A LARGO PLAZO PARA ACELERAR EL INGRESO DE LOS NUEVOS OPERADORES MÓVILES VIRTUALES*. *June*.

Miloslavskaya, N., & Tolstoy, A. (2016). Big Data, Fast Data and Data Lake Concepts. *Procedia Computer Science*, *88*, 300–305. https://doi.org/10.1016/j.procs.2016.07.439

Ministerio de Telecomunicaciones y de la Sociedad de la Información. (2019). *MINTEL Ecuador Digital*. https://www.telecomunicaciones.gob.ec/wp-content/uploads/2019/05/PPT-Estrategia-Ecuador-Digital.pdf

MWC. (2018). *About Mobile World Congress*.

Niño, M., & Illarramendi, A. (2015). Entendiendo El Big Data: Antecedentes, Origen Y Desarrollo Posterior. *Dyna New Technologies*, *2*(3), [8 p.]-[8 p.]. https://doi.org/10.6036/nt7835

Palacio, D. (2020). *Propuesta de una metodología de aseguramiento y control de calidad para los proyectos de software de Inclutec*. *21*(1), 1–9. https://repositoriotec.tec.ac.cr/bitstream/handle/2238/11483/TFG_Dionisio_Palacio.pdf?sequence=1&isAllowed=y

Plasencia Moreno, L., & Anías Calderón, C. (2017). Arquitectura referencial de Big Data para la gestión de las telecomunicaciones. *Ingeniare*, *25*(4), 566–577. https://doi.org/10.4067/S0718-33052017000400566

Salas, R. (2017). Análisis comparativo de herramientas Open Source para Big Data Apache Hadoop y Apache Spark. *Universidad Central Del Ecuador*, *2017*(c), 258. http://www.dspace.uce.edu.ec/bitstream/25000/8059/1/T-UCE-0006-053.pdf%0Ahttp://www.dspace.uce.edu.ec/handle/25000/21351%0Ahttp://www.dspace.uce.edu.ec/handle/25000/20368%0Ahttp://www.dspace.uce.edu.ec/bitstream/25000/12519/1/T-UCE-0015-726.pdf

Sánchez, O. (2019). *Herramientas, retos, oportunidades, suguridad y tendencia del Big Data*. 93. http://ri.uaemex.mx/bitstream/handle/20.500.11799/100155/Tesina.pdf?sequence=3&isAllowed=y

SOLUTIONS, V. (2021). *Radio Access Networks*. 1–8. https://www.viavisolutions.com/es-es/literature/radio-access-networks-interference-analysis-application-notes-en.pdf

The Apache Software Foundation. (n.d.). *Apache Flume*. *0*, 95–107. https://doi.org/10.1007/978-3-319-77800-6_6

Vera Cárdenas, D. J. (2018). *ESTUDIO TÉCNICO PARA LA IMPLEMENTACIÓN DE UNA RED MÓVIL 5G EN LA CIUDAD DE GUAYAQUIL*. http://revistadigital.uce.edu.ec/index.php/CATEDRA/article/download/764/757/.

Yoon, Y., & Kim, W. (2018). Bluff Forwarding: A Practical Protocol for Delivering Refreshed Symmetric Keys on a Multi-Path Big Data Ingestion System. *IEEE Access*, *6*, 24299–24310. https://doi.org/10.1109/ACCESS.2018.2828840

Zheng, K., Yang, Z., Zhang, K., Chatzimisios, P., Yang, K., & Xiang, W. (2016). Big data-driven optimization for mobile networks toward 5G. *IEEE Network*, *30*(1), 44–51. https://doi.org/10.1109/MNET.2016.7389830

ANEXO A

Estructura de un documento XML

La codificación de caracteres es un subconjunto de 8-bit que utiliza el formato de transformación de codificación de caracteres universal (UTF-8 por sus siglas en inglés).

Los caracteres del tipo PrintableString ASN.1 están permitidos, es decir:

- A-Z;
- a-z;
- 0-9;
- <space> ' ( ) + , - . / : = ?'.

En un archivo XML existen los siguientes componentes:

- **Elementos**: Se considera como una pieza lógica del marcado y se representa con una cadena de texto(dato) encerrada entre etiquetas. Pueden existir elementos vacíos (
). Los elementos pueden contener atributos.
- **Instrucciones**: Estas son órdenes especiales para ser utilizadas por la aplicación que procesa

  <?xml-stylesheet type="text/css" href="estilo.css">

  Las instrucciones XML. Comienzan por <? Y terminan por ?>.

- **Comentarios**: Es la información que no forma parte del documento. Comienzan por <!-- y terminan por -->.
- **Declaraciones de tipo**: Especifican información acerca del documento:

  <!DOCTYPE persona SYSTEM "persona.dtd">
- **Secciones CDATA**: Es un conjunto de caracteres que no deben ser interpretados por el procesador:

  <![CDATA[Aquí se puede meter cualquier carácter, como <, &, >, ... Sin que sean interpretados como marcación]]>

La declaración de tipo de documento XML contiene las declaraciones de marcado que proporcionan una gramática para el formato del archivo de medición. Esta gramática se conoce como Definición de tipo de documento (DTD).

```
<!-- MeasDataCollection.dtd version 2.0-->
<?xml version="1.0" encoding="UTF-8"?>
<!ELEMENT mdc (mfh , md*, mff )>
<!ELEMENT mfh (ffv, sn, st, vn, cbt) >
<!ELEMENT md (neid , mi*)>
<!ELEMENT neid (neun, nedn, nesw?)>
<!ELEMENT mi (mts,gp, mt*, mv*)>
<!ELEMENT mv (moid , r*, sf? )>
<!ELEMENT mff (ts)>
<!ELEMENT ts (#PCDATA)>
<!ELEMENT sf (#PCDATA)>
<!ELEMENT r (#PCDATA)>
<!ATTLIST r p CDATA "">
<!ELEMENT mt (#PCDATA)>
<!ATTLIST mt p CDATA "">
<!ELEMENT moid (#PCDATA)>
<!ELEMENT gp (#PCDATA)>
<!ELEMENT mts (#PCDATA)>
<!ELEMENT nedn (#PCDATA)>
<!ELEMENT neun (#PCDATA)>
<!ELEMENT nesw (#PCDATA)>
<!ELEMENT cbt (#PCDATA)>
<!ELEMENT vn (#PCDATA)>
<!ELEMENT st (#PCDATA)>
<!ELEMENT sn (#PCDATA)>
<!ELEMENT ffv (#PCDATA)>

<!-- end of MeasDataCollection.dtd -->
```

El número de etiquetas de resultado de medición (r) por etiquetas de instancia de objeto observado (moid) siempre será igual al número de etiquetas de tipos de medición (mt). En caso de que el resultado sea un valor REAL, el separador decimal será ".". En caso de que el resultado sea "NULL", la marca "r" estará vacía.

El siguiente encabezado se utilizará en archivos de resultados de medición XML reales:

<?xml version="1.0"?>

<?xml-stylesheet type="text/xsl" href="MeasDataCollection.xsl" ?>

<!DOCTYPE mdc SYSTEM "MeasDataCollection.dtd" >

<mdc xmlns:HTML="http://www.w3.org/TR/REC-xml">

- Línea 1: se utilizará la versión xml número 1.

- La referencia a un archivo XSL (Lenguaje de hoja de estilo extensible) o CSS (Hoja de estilo en cascada) en la línea 2 del encabezado es opcional.

- El operador puede configurarlo para que se inserte con el fin de presentar el archivo XML en una GUI de navegador web. Depende del receptor del archivo decidir sobre el uso de esta referencia de hoja de estilo, p. Ej. ignórelo si no es necesario o elija un valor predeterminado configurado si no se proporciona una referencia de hoja de estilo en el archivo.

- Línea 4: Una referencia a la página web de recomendaciones del W3C para XML.

Guía rápida de notación XML:? cero o una ocurrencia

+ una o más ocurrencias

* cero o más ocurrencias

#PCDATA datos de caracteres analizados

ANEXO B

Configuración de la herramienta Kafka

Configuración de Kafka Versión 2.6.0 en Linux Centos

#Iniciar Zookeeper (1er terminal)

sudo bin/zookeeper-server-start.sh config/zookeeper.properties

#Editar el archivo zookeeper

sudo nano config/zookeeper.properties

Iniciar/parar el servicio zookeeper

sudo bin/zookeeper-server-stop.sh

#Configuración de Kafka

sudo nano config/server.properties

#Iniciar Kafka (2do terminal)

sudo bin/kafka-server-start.sh config/server.properties

#Directorio de Logs

/tmp/kafka-logs

#Creación del Topic

sudo bin/kafka-topics.sh --create --bootstrap-server 172.16.9.5:9092 --replication-factor 1 --partitions 1 --topic MyFirstTopic

#Enlistar los Topics

sudo bin/kafka-topics.sh --list --bootstrap-server 172.16.9.5:9092

#Iniciar el Productor (3er terminal)

sudo bin/kafka-console-producer.sh --broker-list 172.16.9.5:9092 --topic MyFirstTopic

#Iniciar el Consumidor (4er terminal)

sudo bin/kafka-console-consumer.sh --bootstrap-server 172.16.9.5:9092 --topic MyFirstTopic --from-beginning

#Importar el archivo XML (Producer) (5to terminal)

sudo cat archivo5G.xml | bin/kafka-console-producer.sh --broker-list 172.16.9.5:9092 --topic MyFirstTopic

Resultados de la configuración obtenidos:

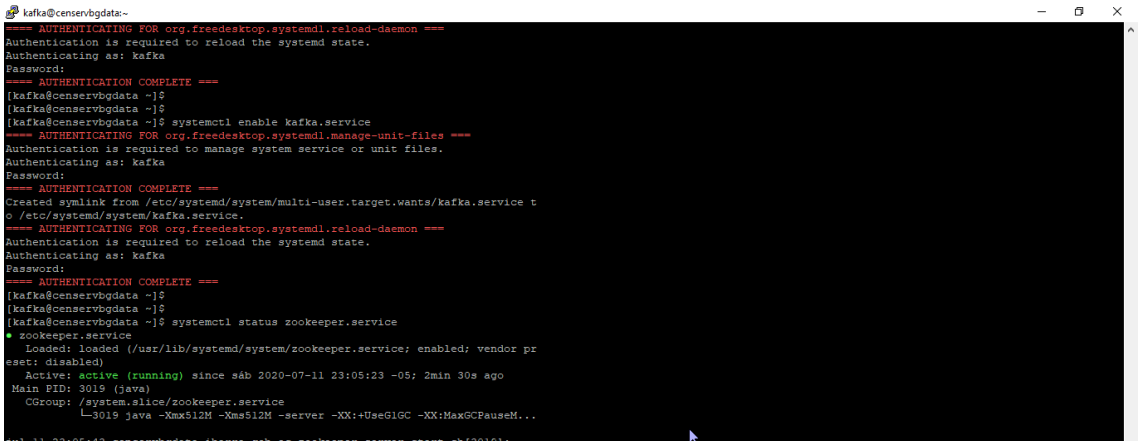


Fig1. Instalación de Kafka

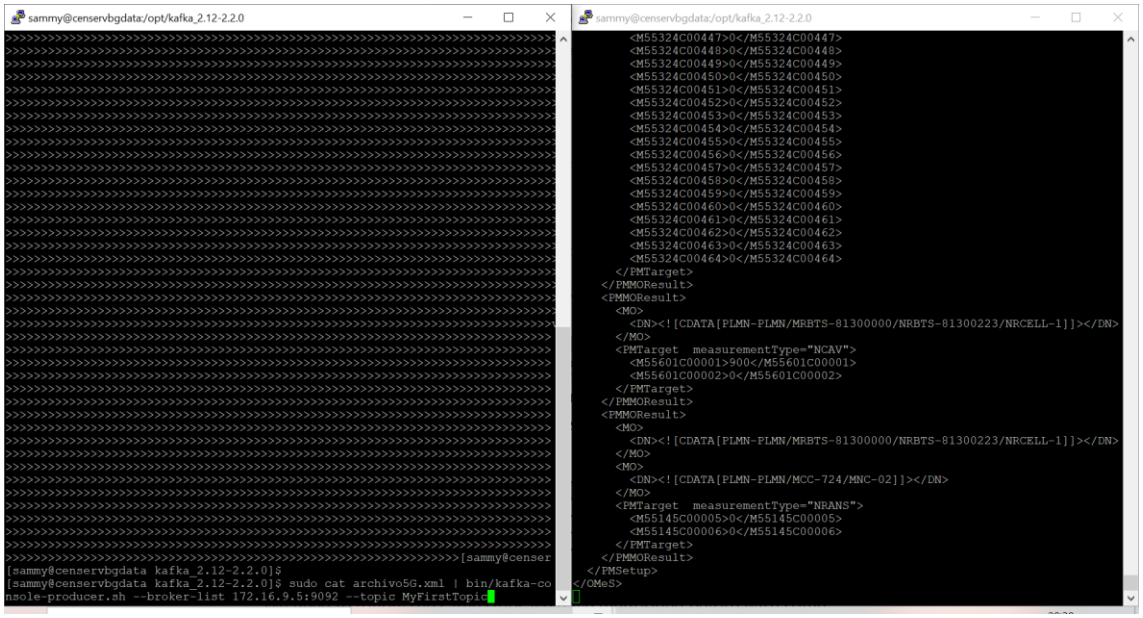


Fig 2. Visualización del archivo XML en HDFS

Configuración de los Conectores de Kafka

Existen varios conectores, para este trabajo se utilizó FileConnector

Configuración del Source Connect

```
curl -X POST \
  /api/kafka-connect/connectors \
  -H 'Content-Type: application/json' \
  -H 'Accept: application/json' \
  -d '{
  "name": "xmlingest",
  "config": {
    "connector.class":
"io.streamthoughts.kafka.connect.filepulse.source.FilePulseSourceConnector",
    "fs.scan.directory.path": "/data/xmls",
    "fs.scan.interval.ms": "10000",
    "fs.scan.filters":
"io.streamthoughts.kafka.connect.filepulse.scanner.local.filter.RegexFileListFilter",
    "file.filter.regex.pattern": ".*\\.xml$",
```

```
    "task.reader.class":
"io.streamthoughts.kafka.connect.filepulse.reader.BytesArrayInputReader",
    "force.array.on.fields": "track",
    "offset.strategy": "name",
    "topic": "xmlreader-load",
    "internal.kafka.reporter.bootstrap.servers": "192.168.100.16:9092",
    "internal.kafka.reporter.topic": "connect-file-pulse-status",
    "fs.cleanup.policy.class":
"io.streamthoughts.kafka.connect.filepulse.clean.LogCleanupPolicy",
    "tasks.max": 1,
    "value.converter": "org.apache.kafka.connect.json.JsonConverter",
    "value.converter.schemas.enable": "false"
  }
}' http://127.0.0.1:8083/connectors
```

Configuración del Sink Connect

```
curl -X POST \
  /api/kafka-connect/connectors \
  -H 'Content-Type: application/json' \
  -H 'Accept: application/json' \
  -d '{
  "name": "hdfs-sink22212",
  "config": {
    "connector.class": "io.confluent.connect.hdfs.HdfsSinkConnector",
    "topics": "xmlreader-load-01",
    "tasks.max": "1",
    "flush.size": "3",
    "hdfs.url": "hdfs://192.168.100.16:9000",
    "key.converter": "org.apache.kafka.connect.json.JsonConverter",
    "value.converter": "org.apache.kafka.connect.json.JsonConverter",
```

```
    "key.converter.schemas.enable": "false",
    "value.converter.schemas.enable": "false",
    "store.url": "hdfs://192.168.100.16:9000",
    "format.class": "io.confluent.connect.hdfs.json.JsonFormat"
  }
}' http://127.0.0.1:8083/connectors
```

ANEXO C

Configuración de las herramientas Flume y Hadoop

Configuración de Flume Versión 1.9.0

```
# Sources, channels, and sinks are defined per
# agent name, in this case 'tier1'.
tier1.sources  = source1
tier1.channels = channel1
tier1.sinks    = sink1

# For each source, channel, and sink, set
# standard properties.
# source details
tier1.sources.source1.type     = spooldir
tier1.sources.source1.spoolDir                                                    =
/home/eduwer/Documentos/practicas/xmlingest/xmlingest/muestra14
tier1.sources.source1.fileHeader = false
tier1.sources.source1.basenameHeader = true
tier1.sources.source1.fileSuffix  = .COMPLETED3
tier1.sources.source1.thread = 1
tier1.sources.source1.channels = channel1
tier1.sources.source1.fileHeader = true
tier1.sources.source1.basenameHeader =true
tier1.sources.source1.batchSize = 1000

# channel details
tier1.channels.channel1.type   = memory
tier1.channels.channel1.capacity = 60000000
tier1.channels.channel1.transactionCapacity = 50000000
```

tier1.channels.channel1.byteCapacityBufferPercentage = 20

tier1.channels.channel1.byteCapacity = 6442450944

sink details

tier1.sinks.sink1.type = HDFS

tier1.sinks.sink1.fileType = DataStream

tier1.sinks.sink1.channel = channel1

tier1.sinks.sink1.hdfs.path = hdfs://localhost:9000/datosmuestra14

tier1.sinks.sink1.hdfs.filePrefix = %{basename}

tier1.sinks.sink1.hdfs.batchSize = 1000

tier1.sinks.sink1.hdfs.rollInterval = 0

tier1.sinks.sink1.hdfs.rollSize = 0

tier1.sinks.sink1.hdfs.rollCount = 0

tier1.sinks.sink1.hdfs.idleTimeout = 60

Configuración de Hadoop Versión 3.2.2

Para direccionar los datos desde el administrador del elemento hacia Hadoop, este comando se lo pone en Hadoop y se va cambiando la muestra para cada prueba.

hadoop fs -copyFromLocal ./muestra10 /muestra10