UNIVERSIDAD ANDRES BELLO

**FACULTAD DE INGENIERÍA**

**INGENIERÍA EN COMPUTACIÓN E INFORMÁTICA**



Proyecto Big Data

Seminario de Título 1

**David Córdova Cordero**

**VIÑA DEL MAR – CHILE**

**Mayo 2019**

Contenidos

[Introducción 8](#_Toc13006947)

[Definición de Big data 9](#_Toc13006948)

[Atributos de Big Data 9](#_Toc13006949)

[Análisis de la problemática. 12](#_Toc13006950)

[Técnica de los cinco por qué. 12](#_Toc13006951)

[Diagrama Ishikawa 13](#_Toc13006952)

[Objetivo del proyecto 14](#_Toc13006953)

[Descripción del problema 14](#_Toc13006954)

[Objetivo del proyecto 15](#_Toc13006955)

[Objetivo general 15](#_Toc13006956)

[Objetivos específicos 15](#_Toc13006957)

[Diagrama de alto nivel 17](#_Toc13006958)

[Arquitectura de solución 17](#_Toc13006959)

[Solución ideal 1](#_Toc13006960)

[Limitaciones 1](#_Toc13006961)

[Restricciones 2](#_Toc13006962)

[Alternativas de solución 2](#_Toc13006963)

[Procedimientos manuales 2](#_Toc13006964)

[Cambios en procedimientos actuales 2](#_Toc13006965)

[Alternativas disponibles en el mercado. 3](#_Toc13006966)

[Solución propuesta 3](#_Toc13006967)

[Situación Futura 3](#_Toc13006968)

[Factibilidad operativa 4](#_Toc13006969)

[Factibilidad técnica 4](#_Toc13006970)

[Factibilidad Económica 5](#_Toc13006971)

[Historias de Usuario 6](#_Toc13006972)

[Requerimientos de alto nivel 6](#_Toc13006973)

[Requerimientos Sistema: 6](#_Toc13006974)

[Requerimientos Funcionales: 7](#_Toc13006975)

[Requerimientos No Funcionales: 7](#_Toc13006976)

[Delimitación 8](#_Toc13006977)

[El desarrollo del proyecto incluye: 8](#_Toc13006978)

[Plan de proyecto 9](#_Toc13006979)

[Dirección de proyecto 9](#_Toc13006980)

[Tabla de roles 11](#_Toc13006981)

[Posteriormente, se desarrollan varios documentos: 11](#_Toc13006982)

[Scrum 11](#_Toc13006983)

[Product backlog del sprint 12](#_Toc13006984)

[Arquitectura de un ambiente de big data 13](#_Toc13006985)

[Arquitectura de Big Data 13](#_Toc13006986)

[Arquitectura de un ambiente de Big data 14](#_Toc13006987)

[14](#_Toc13006988)

[Ingreso de datos 14](#_Toc13006989)

[Gestión de datos 15](#_Toc13006990)

[Tiempo real de procesamiento 15](#_Toc13006991)

[Análisis de datos 15](#_Toc13006992)

[Bases de datos NoSQL 16](#_Toc13006993)

[Distribuido 16](#_Toc13006994)

[Escalabilidad horizontal 16](#_Toc13006995)

[Construido para grandes volúmenes 16](#_Toc13006996)

[Modelos de datos no relacionales 16](#_Toc13006997)

[No hay definiciones de esquema 17](#_Toc13006998)

[Hadoop y big data 17](#_Toc13006999)

[Ventajas 17](#_Toc13007000)

[Arquitectura hadoop 17](#_Toc13007001)

[Hadoop 18](#_Toc13007002)

[Arquitectura hadoop 18](#_Toc13007003)

[MapReduce 18](#_Toc13007004)

[Función Map 18](#_Toc13007005)

[Función Reduce 19](#_Toc13007006)

[Proceso MapReduce 19](#_Toc13007007)

[HDFS 19](#_Toc13007008)

[Chukwa 20](#_Toc13007009)

[Diagrama de Chukwa 20](#_Toc13007010)

[21](#_Toc13007011)

[Sqoop 21](#_Toc13007012)

[Flume 21](#_Toc13007013)

[Flujo de datos 22](#_Toc13007014)

[Avro 22](#_Toc13007015)

[Hive 22](#_Toc13007016)

[Pig 23](#_Toc13007017)

[Pig posee las siguientes características: 23](#_Toc13007018)

[Facilidad de programación 23](#_Toc13007019)

[Oportunidades de optimización 24](#_Toc13007020)

[Extensibilidad 24](#_Toc13007021)

[Para la implementación de las herramientas 24](#_Toc13007022)

[Metodología iterativo incremental 24](#_Toc13007023)

[Ciclo de vida iterativo 24](#_Toc13007024)

[Ciclo de vida incremental 25](#_Toc13007025)

[Modelo de desarrollo 25](#_Toc13007026)

[Modelo Iterativo y acoplamiento al proyecto 25](#_Toc13007027)

[Ventaja 26](#_Toc13007028)

[Desventajas 26](#_Toc13007029)

[Requerimientos de Hardware y software 26](#_Toc13007030)

Resumen

En el presente documento se tiene como meta entregar una propuesta de solución a la problemática que presentan todas aquellas empresas que manejen grandes volúmenes de datos.

Hoy en día gran cantidad de empresas de empresas lo largo de todo chile tiene el problema del mal uso de sus datos ya sea por la escasez de tecnología en sus sistemas, por desconocimiento de las herramientas en el mercado y/o por el uso inadecuado de herramientas.

La materia de los datos necesita ser procesados para ser convertidos en información importante para que ayude a tomar decisiones dentro de las empresas y es lo que hoy en día se presenta como una escasez o como falta de información al respecto de estas tecnologías.

Dado a lo anterior, se propone realizar una investigación respecto a las plataformas que soporten estos grandes volúmenes de datos llamados Big Data, y dar conocer el funcionamiento de estos, lo importantes que puede llegar a ser la buena elección de unas de estas herramientas al interior de las empresas y por ultimó como ayudan en la toma de decisiones para resolver grandes problemas.

CAPÍTULO 1: INTRODUCCIÓN

# Introducción

Con el constante crecimiento de información en cada uno de los aspectos más relevantes mundialmente como sociedad, comercio y ciencia, se vuelve necesario un cambio con respecto al manejo de la información, que hasta hace poco tiempo se venía implementando a partir de bases de datos relacionales. Si bien este esquema funcionaba para ambientes que almacenaban teras de información con datos puntuales y que a través de consultas SQL se podía buscar la información requerida, se vuelve dispendioso y costoso al momento de analizar un mayor volumen de información, en ambientes donde anualmente se almacenan petas de todo tipo de datos, incluyendo aquí los no estructurados, como archivos de video y audio; un ejemplo de esto es “Google, que recibe a diario trillones de bytes, con el objetivo de ofrecer muchos de los servicios que actualmente se conocen como el mismo motor de búsqueda y Google Earth.

El manejo de grandes cantidades de información conlleva a pensar en la implementación de herramientas que permitan administrar y gestionar este tipo de datos no estructurados y semiestructurados, en la búsqueda de patrones concurrentes para la toma de decisiones. Actualmente se pueden encontrar tecnologías como Hadoop, MapReduce y bases de datos NoSQL, que se pueden implementar en la creación de un ambiente Big Data.

De esta manera, el proyecto busco la implementación de una arquitectura para crear un ambiente Big Data en la Universidad Andrés Bello de viña del mar, teniendo en cuenta aspectos importantes como el software y el hardware que se debía utilizar para realizar dicha labor, de igual manera todos los procedimientos que implicaba empezar a utilizar bases de datos no relacionales. Todo esto se ve reflejado a través de dos guías, donde se podrá encontrar de una manera detallada los pasos para la implementación de un ambiente Big Data, sus elementos y procesos para ingresar y consultar información a través de un ejemplo específico, con el propósito de diseñar estrategias y buscar patrones que permitan una buena gestión de la información.

# Definición de Big data

Big data es una referencia a aquellos sistemas de información que manejan conjuntos de datos de gran volumen, de alta velocidad, de veracidad, de valor y de gran variedad de recursos, que demandan formas rentables e innovadoras de procesamiento de la información para mejorar la comprensión y la toma de decisiones.

Big data es la solución al crecimiento exponencial de los datos, en el momento en que se hace difícil su administración con respecto al almacenamiento, procesamiento y acceso.

De esto se puede obtener beneficios como:

* Optimizar el cálculo y la precisión algorítmica para reunir, analizar, enlazar y comparar conjuntos de grandes datos.
* Identificar patrones para la toma de decisiones en los ámbitos económico, social, técnico y legal.

La mayoría de las definiciones que se pueden encontrar de Big data están enfocadas al volumen de los datos, al almacenamiento de dicha información, de esto se puede concluir que el volumen importa pero que también existen otros atributos importantes de Big data, estos son: La velocidad, la veracidad, la variedad y el valor. Estos cinco aspectos constituyen una definición comprensiva y además destruyen el mito acerca de que Big data se trata únicamente del volumen. A cada uno de estos aspectos se le atribuyen las siguientes características:

# Atributos de Big Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Volumen** | **Velocidad** | **Variedad** | **Veracidad** | **Valor** |
| Almacenamiento En terabytes | Por lotes | Estructurado | Integridad y Autenticidad | Estadísticas |
| Registros | Tiempo Cercano | No estructurado | Origen y  Reputación | Eventos |
| Transacciones | Tiempo Real | Multi-factor | Disponibilidad | Correlaciones |
| Tablas y  Archivos | Procesos | Probabilística | Responsabilidad | Hipótesis |

Tabla : 5 Vs

Con el paso del tiempo, las empresas han fomentado la creación de nuestras estrategias para la toma de decisiones, dando un importante lugar al análisis predictivo, ya que con esto se han podido determinar diversos tipos de patrones entre la sociedad, generando como consecuencia gran cantidad de beneficios consistentes en la innovación, investigación y desarrollo de nuevas soluciones.

La generación de estos datos en los últimos años ha venido creciendo de manera inmensurable y se proyecta a seguirlo, por tal motivo es que Big data se convertirá en uno de los principales aspectos a tener en cuenta dentro de los ámbitos comerciales, científicos y social, todo debido al impacto económico e innovador que este ha representado.

Lo anterior se puede ver reflejado en una recopilación de estadísticas nombrada “A comprehensive list of Big data statistics”, donde se exponen algunos puntos del por qué la importancia de Big data, algunos de ellos son:

* Actualmente en el mundo digital existe 2.7 zeta bytes de datos.
* El gobierno de estadounidense invierte cerca de 200 millones de dólares en investigación sobre Big data.
* La red social Facebook almacena, registra y analiza diariamente 30 pata bytes de datos, 94% de los usuarios de Hadoop realiza análisis de grandes volúmenes de información que antes no se podía analizar

Descifrar el genoma humano tardo cerca de 10 años, actualmente ese proceso se puede realizar en una semana.

CAPÍTULO 2: FUNDAMENTACIÓN

# Análisis de la problemática.

## Técnica de los cinco por qué.

Por qué implementar una plataforma de soporte proyectos Big Data

1. ¿Por qué los datos de las empresas son ilocalizables e inutilizables?

Por que manejan demasiado volumen de datos y no utilizan las herramientas y plataformas correctas acorde a los datos que maneja la empresa

1. ¿Por qué que es importante usar las herramientas y plataformas adecuadas para el buen uso de los datos?

Por qué las empresas manejan distintos tipos de datos, los que son; estructurados, semi estructurados y no estructurados y deben tener las herramientas necesarias para el manejo de estos tipos de datos para que así se comuniquen de la mejor forma con las plataformas.

1. ¿Por qué utilizar plataformas que den soporte a proyecto big data?

Porque dependen de que datos use la empresa dependerá de que tecnología se use para desarrollar una plataforma que soporte los datos que maneja la empresa ya sean datos estructurado como no estructurados.

1. ¿Por qué una plataforma sería la solución a la entrega de información tiempo real?

Por qué una plataforma que trabaje con la tecnología adecuada podrá brindar y almacenar datos como el usuario estime conveniente y así entregarle información a tiempo real.

1. ¿Por qué Big data suele ser un problema para las empresas que no utilizan herramientas para interactuar con estas?

Por qué las Big datas son inmensos volúmenes de datos que al momento de solicitarlo las personas se demora un tiempo considerable en localizar los datos para poder ocuparlos, si se ocupa una herramienta que trabaje con Big data y una plataforma que interactúe con esta herramienta se podría localizar los datos acortando el tiempo de localización de los datos para poder analizarlos y poder ocuparlos como información esencial para la empresa.

La utilización de herramientas para manejar los datos y que interactúen con las plataformas de soporte de Big data es un avance considerable hacia el tiempo de respuesta que una empresa debe tener para las buenas tomas de decisiones y para corregir errores considerables dentro de esta.

# Diagrama Ishikawa

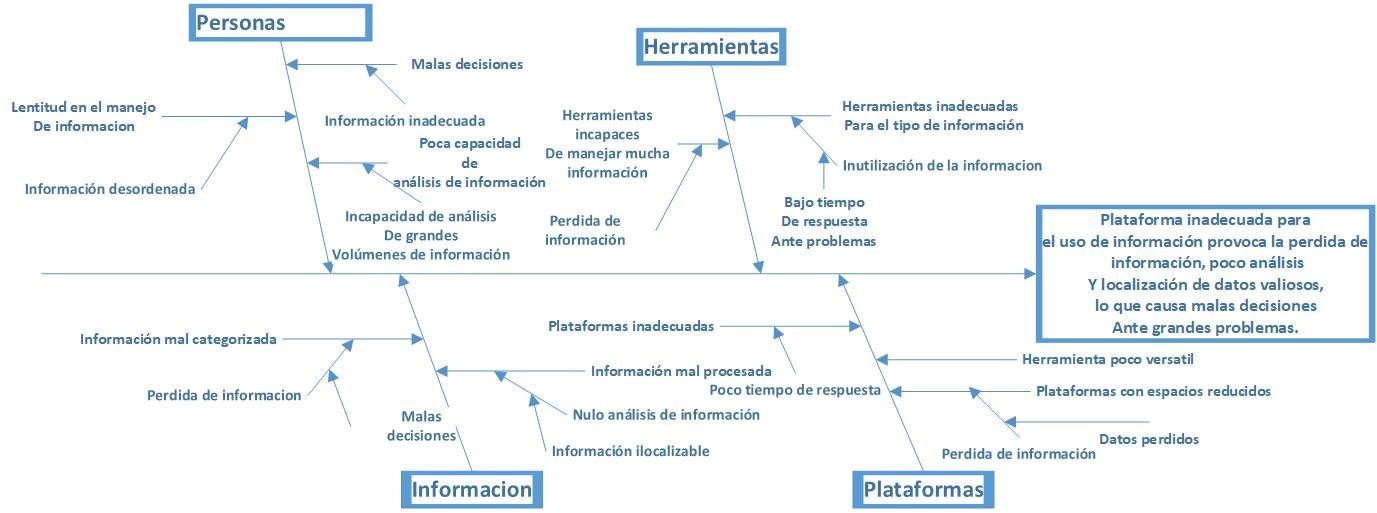


Ilustración : Diagrama de Ishikawa

En el diagrama de Ishikawa se presentan las principales causas de porque hay un rechazo a las plataformas existentes. La mala elección de una plataformas para trabajar los datos puede provocar la mala gestión de estos, provocando que estos se conviertan en información relevante para los encargados de manejarlos , causando mala toma de decisiones.

# Objetivo del proyecto

## Descripción del problema

Big data es un concepto nuevo que se está desarrollando hace pocos años y que solo las grandes empresas han iniciado a implementarlo, la documentación que se encuentra sobre este tema es poca y la mayoría se encuentra en distintos idiomas, especialmente en inglés, las implementación que se han hecho son aisladas y hay pocas guías paso a paso que permita dar soluciones a los errores y fallos que se presentan y que permita realizar la integración de las herramientas open source que se puedan encontrar.

Existen organizaciones y empresas que no pueden pagar o tener una consultoría adecuada sobre el manejo de las nuevas tendencias Big Data, que les permita seleccionar las herramientas adecuadas y la forma de configurar un ambiente para adaptarlo a su modelo de negocio.

Big data es la combinación de viejas y nuevas tecnologías que ayudan a las empresas a obtener una mejor visión del procesamiento de sus informaciones. También puede conceptualizar como la capacidad de manejar un inmenso volumen de datos que se generan de forma caótica, que, a la velocidad y temporización correctas, permite el análisis a tiempo real y la definición de las acciones asociadas necesarias.

Los datos no funcionan y no se les está sacando el mayor partido porque no se están usando bien debido a que los datos no están en el lugar donde deben estar. Según sus estimaciones de experto el 80% de los datos de las empresas no están almacenados de buena forma o no están siendo utilizados por las herramientas y plataformas adecuada lo que provoca la pérdida de información y pérdida de tiempo en dar respuestas a problemas de las empresas.

Según expertos solo el 54 % de las empresas hoy en día tienen localizados sus datos, mientras que el restante porcentaje tiene el gran problema de pérdidas de datos.

## Objetivo del proyecto

El objetivo de este proyecto es dar a conocer estas grandes herramientas tecnologías y brindar apoyo a las plataformas ya existentes de manejo de datos. Hay una enorme cantidad de datos e los que queremos procesar y extraer información, en donde no se puede utilizar metodologías tradicionales ya que todas las imágenes, bases de datos, correos, Facebook, Twitter, sensores, etc. Son información estructurada como no estructurada es decir información que podemos extraer de datos totalmente dispares o bien de nuestra base de datos tradicional todo esto lo procesaremos y lo meteremos dentro de un entorno big data, para poder procesar esa información y extraer la información que necesitamos, dado que las tecnologías tradicionales no pueden hacer frete a esta cantidad de información es necesario utilizar nuevas tecnólogas, es por lo siguiente que se utilizara informática distribuida, se distribuirán datos y procesos que es en lo que se basa Big data hadoop.

## Objetivo general

Diseñar y desarrollar una guía para la implementación de un ambiente Big data en la universidad Andrés Bello, la cual ayude a la instalación y a la puesta en marcha de la herramienta hadoop y sus herramientas derivadas open source, la cual permita el manejo de altos volúmenes de datos para así dar a conocer y enseñar el manejo de este ambiente, para incremental en al menos un 60% su manejo.

## Objetivos específicos

* Identificar la tecnología, herramientas de software y requerimientos de hardware necesarios para la implementación de un ambiente de Big data.
* Realizar el montaje, configuración e integración de las herramientas para la construcción del ambiente de Big data
* Diseñar el caso de estudio que se utilizara para realizar las pruebas del ambiente configurado
* Implementar la herramienta apache hadoop y las herramientas derivadas de este para el uso de datos.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Objetivo específicos | Situación Actual | Objetivo especifico | Situación esperada | Métricas | Criterio de éxito |
| 1 | Desconocimiento de herramientas y software Open Source que trabajen en conjunto con hadoop. | Identificar la tecnología, herramientas de software y requerimientos de hardware necesarios para la implementación de un ambiente de Big data. | Generar un documento de memoria, el cual contenga la información del ambiente big data. | Documentos con información relevantes | Documento con información relevante >= 100% |
| 2 | Manejo de volúmenes de información de forma manual en discos | Realizar el montaje, configuración e integración de las herramientas para la construcción del ambiente de Big data | Realizar la elevación de los servicios de las herramientas hadoop | Elevación de servicios – no elevación de los servicios | Elevación de servicios =5 |
| 3 | Sin casos de estudios | Diseñar el caso de estudio que se utilizara para realizar las pruebas del ambiente configurado | Realizar casos de estudios para las pruebas de ambiente | Casos de estudios | >=60% de casos de estudios |
| 4 | Sin implementación de las herramientas | Implementar la herramienta apache hadoop y las herramientas derivadas de este para el uso de datos. | Montaje de la herramienta hadoop con las herramientas apache necesarias para el manejo de datos. | N° de servicios | >=3 servicios |

Tabla 2: Métricas de Objetivos específicos

La importancia de desarrollar el proyecto radica en apoyar el aprendizaje de las nuevas tendencias en la gestión de la información.

Así mismo el proyecto contribuye y apoya al análisis más detallado a los distintos casos que se puedan presentar y que no puedan ser resueltos mediante bases de datos relacionales.

## Diagrama de alto nivel

A continuación, se presenta el diagrama de alto nivel de la solución para la plataforma de soporte de proyecto Big data.

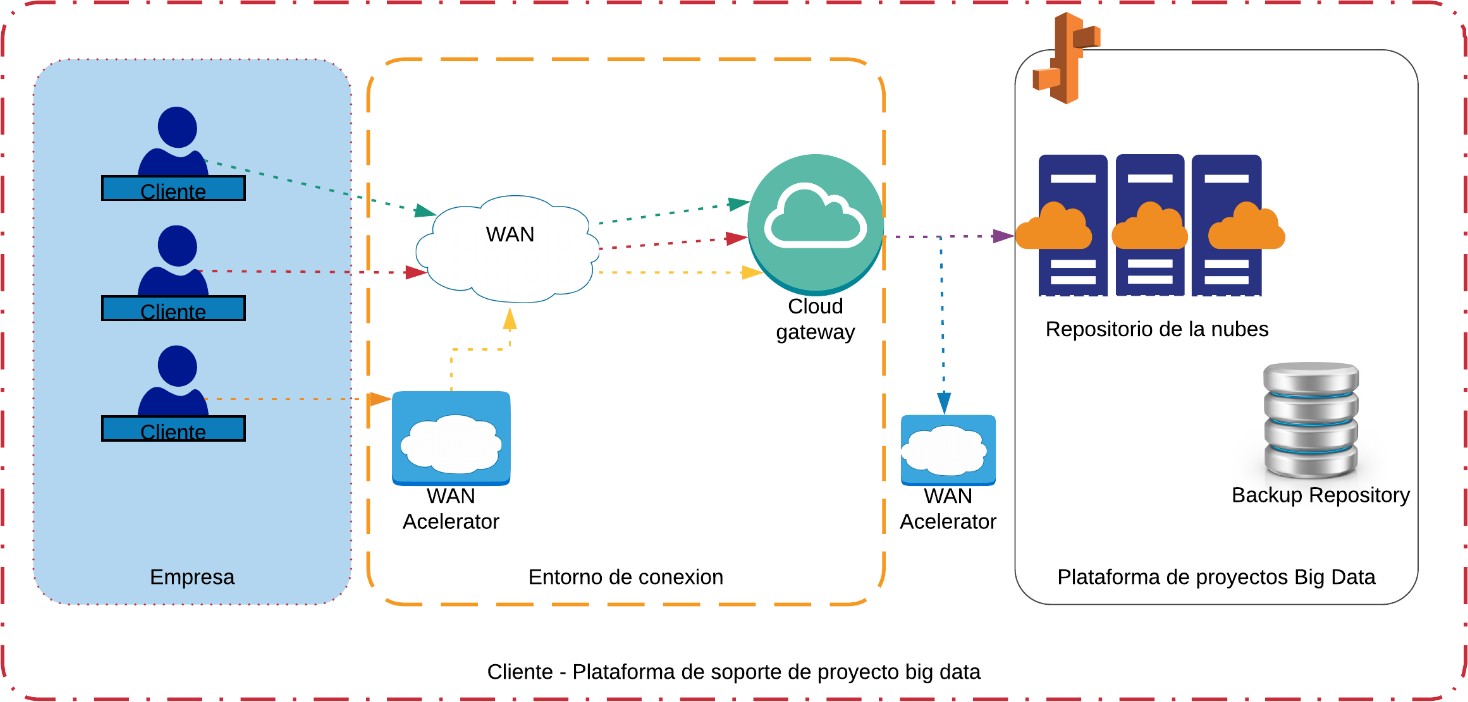


Ilustración 2: Diagrama de alto nivel

## Arquitectura de solución

El sistema es capaz de captar información de fuentes externas con el fin de convertirla en datos de valor mediante técnicas de Procesado del Lenguaje Natural en combinación con tecnologías de procesado y almacenamiento Big Data*.* La información utilizada es obtenida de redes sociales, canales RSS, bases de datos de patentes, artículos científicos, etc, donde los usuarios alimentan al sistema de conocimiento, añadiendo de forma continua nuevas fuentes. De esta manera, se crea una base de conocimiento que escala en dimensiones y que proporciona información para poder conocer el mercado y los competidores próximos a nuestro entorno de una forma automatizada.

La arquitectura está formada por un sistema Cloud Computing compuesto por tres capas: infraestructura como servicio (IaaS), plataforma como servicio (PaaS) y software como servicio (SaaS)

* La primera capa del modelo cloud, Infraestructura como servicio (IaaS), proporciona una gestión bajo demanda de todos los recursos de hardware disponibles en un clúster, lo que facilita, mediante técnicas de virtualización, una rápida disponibilidad de los recursos necesarios para las tareas de computación.
* En la capa de Plataforma como Servicio (PaaS) se ofrece un framework de desarrollo formado por un ecosistema Big Data compuesto por diversas herramientas basadas en el paradigma mapreduce , que proporcionan la capacidad de procesar grandes cantidades de información en paralelo.

La última capa del modelo cloud, Software as a Service (SaaS), proporciona al usuario final una aplicación web donde se ofrece toda la información que ha sido generada por el sistema de forma gráfica y sintetizada. La información también se puede obtener mediante interfaces de ejecución e interfaces de consulta a través de un API REST para poder ofrecer información de valor desde otro tipo de dispositivos o aplicaciones de terceros.

La adquisición de información se realiza mediante capaz modulares de extracción, transformación y cargas (ETL), encargados de capturar información de diferentes recursos (redes sociales, páginas web, documentos corporativos y científicos, patentes). La arquitectura está preparada para la captura de información, tanto estructurada como no estructurada, donde herramientas del ecosistema hadoop juegan un papel fundamental tanto en fase de análisis y procesamiento del lenguaje natural como en fase de predicción y clasificación.

* + Apache Pig: Plataforma de alto nivel para crear programas MapReduce utilizados en

Hadoop

* + Apache hive: Infraestructura de almacenamiento de datos construidas sobre hadoop para proporcionar agrupación, consulta, y análisis de datos.
  + Apache mahout: Es un proyecto para producir implementaciones gratuitas de algoritmo de aprendizajes automáticos distribuidos enfocados principalmente en las áreas de filtrado. Mahout también proporciona bibliotecas de java para operaciones matemáticas comunes enfocadas en algebra lineal y estadística

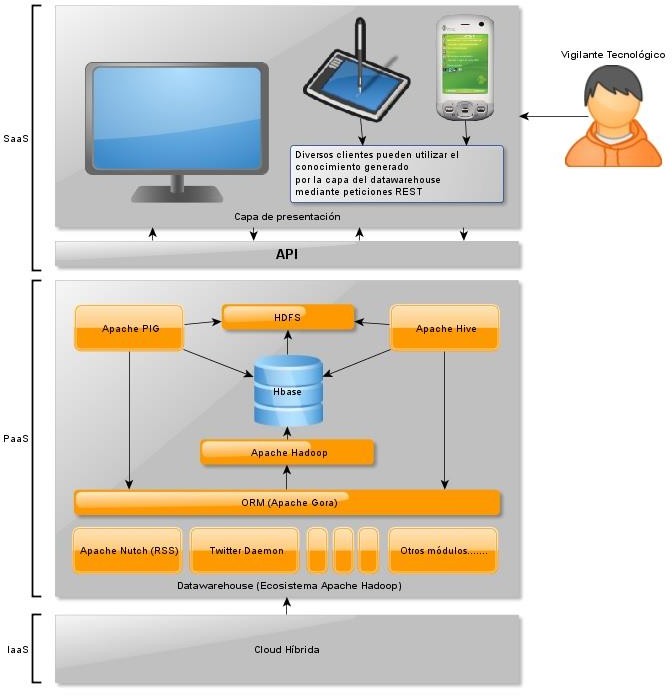


Ilustración 3: Arquitectura de solución hadoop

# 

## Solución ideal

El alcance de la solución propone generar una guía para la implementación del ambiente big data, para así realizar la creación de una nueva plataforma open source que permita el manejo de datos y pueda mostrar a las personas como funciona el ambiente, esto permitirá tener un ambiente de pruebas para esta herramienta y observar las mejoras que trae su implementación.

## Limitaciones

* Los usuarios deben tener acceso a internet para ocupar la plataforma.
* Deben utilizar un browser o explorador web para acceder a la plataforma.
* No envía notificaciones
* Los datos perdidos en el equipo no es responsabilidad de la plataforma.

## Restricciones

* La plataforma deberá mantener a la vista los datos almacenados.
* Los usuarios tendrán una plataforma sin diseño gráfico.
* La plataforma solo muestra como son manejados los datos.

# Alternativas de solución

## Procedimientos manuales

Como primera alternativa, se considera continuar con el proceso manual, disminuir el uso de datos y aumentar los usuarios a cargos de los datos para que estos sean localizados.

Para el análisis de información, categorizar los datos, dando mayor prioridad a los datos no estructurados, para así manejarlos con mayor tiempo y poder hacer uso de ellos.

## Cambios en procedimientos actuales

Si bien es cierto existen plataformas que trabajan con las Big datas, muchas de las plataformas no dan a conocer por que es importante el usarlas y cómo utilizarlas. Es de suma importancia el mostrar cómo se utilizan las nuevas plataformas para que los futuros clientes confíen en ellas y observen los buenos cambios que generan.

Se debe implementar una guía para dar a conocer las instalaciones del ambiente hadoop y sus herramientas para el manejo de datos, para dar a conocer cómo trabajan estas herramientas en conjunto, para así implementar un ambiente de pruebas, esto permitirá que los usuarios puedan usar de mejor forma estos servicios y lograr entender lo importante que es el uso de estas herramientas en el manejo de datos.

## Alternativas disponibles en el mercado.

Algunas de las alternativas disponibles en el mercado y que entregan una plataforma con una gran variedad de sistemas son: Amazon web Service y Microsoft Azure, que son dos de las grandes plataformas que se dedican a prestar estos servicios de soporte de proyecto big data.

Amazon web engloba una gran cantidad de servicios para poder realizar distintos tipos de actividades en la nube. Desde almacenamiento a la gestión de instancias, imágenes virtuales, desarrollo de aplicaciones móviles, etc.

Microsoft Azure es una nube publica de pago por uso que permite compilar, implementar y administrar rápidamente aplicaciones en una red global de datacenters de Microsoft.

Estos servicios ya importantes en el mercado trabajan con herramientas hadoop y sus servicios derivados, solo muestran al usuario un entorno de diseño para que ellos trabajen con sus datos, pero no muestra cómo pueden crear y usar en su completitud estas herramientas.

## Solución propuesta

Si bien las aplicaciones que dan soporte a los grandes volúmenes de datos ya son bastante completas, y la mejora que se puede implementar son bastantes reducidas, pero aun así existen. Estas pueden ser:

Realizar e implementar una guía para la instalación de hadoop en el ambiente big data, con ello viene la incorporación de la instalación del ambiente hadoop y sus servicios derivados, para así dar un entorno de pruebas y manejo de datos dentro de la universidad Andrés bello.

Ofrecer multiplataforma seguras monitoreadas y rastreables.

## Situación Futura

Se debe realizar una investigación en profundidad para saber que herramientas le darán un mayor potencial a la plataforma, esto permitirá extender la guía de implementada para el ambiente haddoop, para tener un entorno más extenso con servicios open source.

En la situación futura lo ideal es la combinación de distintas herramientas que operen en conjunto para entregar al usuario cliente una plataforma rápida, segura y amigable a la vista del cliente para que este al momento de dar uso de ella pueda sacar todo el provecho y tomas las mejores decisiones antes errores en el menor tiempo posible.

## Factibilidad operativa

La factibilidad operativa comprende una determinación de la probabilidad de que un nuevo sistema se use como se supone, aprovechando los beneficios que entrega a todos los usuarios clientes.

* + Operación garantizada
  + Uso garantizado

El correcto funcionamiento de la plataforma dependerá siempre a la capacidad de los empleados para poder utilizarlo, la utilización de herramientas que entreguen apoyo en las labores de trabajo permite encontrar los fallos y dar a conocer las necesidades de una forma más fácil y amigable. Basándose en las cuantiosas investigaciones realizadas y la presencia de plataformas que ya realizan este trabajo es factible operacionalmente realizar la plataforma o implementar mejoras en las ya existentes.

## Factibilidad técnica

La factibilidad técnica consiste en realizar una evaluación de las herramientas tecnológicas que dan soporte a la plataforma Big data y la posibilidad de usar las mismas herramientas para mejorar la plataforma y/o combinar nuevas herramientas que funcionen en conjunto con las antiguas.

Hoy en día las plataformas más grandes tanto Amazon web service como Microsoft Azure ocupan herramientas distintas para poder realizar sus labores, pero con las investigaciones realizadas se corrobora que ambas empresas si pueden implementar herramientas que la competencia ocupa para combinarlas con las que utilizan, todo esto para dar mejora a sus sistemas, u/o realizar nuevas plataformas para abarcar más soluciones, ya que ambas Ocupan repositorios en la nube, tienen conexiones similares y ambas trabajan de la misma forma, solo se debe tener en cuenta al momento de elegir la herramienta que tipo de datos soporta y maneja. Dado lo comentado anteriormente es factible técnicamente.

## Factibilidad Económica

La Factibilidad Económica es de vital importancia, dado que la utilización de nuevas herramientas conlleva otros costos los cuales se deben solventar, como la adquisición de servidores o alojamientos de datos en la nube es por eso que se debe tener claro si se tienen los recursos para poder financiar el uso de nuevas herramientas y nubes para albergar la plataforma.

En el siguiente cuadro se podrá visualizar la comparación entre las 2 plataforma hoy más importante en el manejo de datos y la plataforma hadoop implementada en el proyecto.

|  |  |  |  |
| --- | --- | --- | --- |
| **Factibilidades** | **Microsoft Asure** | **Amazon web Service** | **Proyecto Apache Hadoop** |
| **Operativa** | -Plataforma con múltiples funciones para el manejo de datos  -Sistema con soporte técnico ante fallas  -Plataforma sin manual de uso  -Plataforma con información completa respecto a los servicios ofrecidos  -Tecnología de aprendizaje automático | -Plataforma con múltiples funciones para el manejo de datos y otros servicios  -Sistema con soporte técnico  -Plataforma sin manual de uso de sus servicios  -Plataforma con información completa respecto a los servicios que entrega  -Tecnología de aprendizaje automático | -Plataforma con servicios para el manejo de datos  -Soporte ante fallas  -Manual de uso de las herramientas  -Manual de instalación de las herramientas  -Ambiente de pruebas de herramientas apache |
| **Técnica** | -Data Lake Analytics  -HDInsight  -Azure Stream Analytics  -Azure es Data Lake Store  -Hadoop  -StorSimple  ­-Power BI | -Amazon Elastic MapReduce  -Hadoop  -Apache Spark  -Kinesis Firehose  -Kinesis Streams  -Redshift  -Amazon Elasticsearch  -DynamoDB | -Hadoop  -Apache hive  -Apache habase  -Apache pig  -Apache Spark |
| **Económica** | -Plan Developer 27,47 USD  -Plan Standard 94,71 USD  -Plan Professional Direct 947,07 USD  -Plan Premier Sin information | -Plan Basic Free  -Plan Developer 60 USD  -Plan Business 6150 USD  -Plan Enterprise 70500 USD | -Open source FREE |

# Historias de Usuario

* HU1.- Investigar sobre una plataforma y herramienta que trabajen en conjunto y que permiten que múltiples usuarios puedan estar conectados al mismo tiempo para realizar distintas operaciones en una misma instancia de tiempo
* HU2.- Dar a conocer cómo se acopla Big data con las herramientas hadoop
* HU3.- Mostrar el funcionamiento exitoso de la herramienta hadoop
* HU4.- Realizar las pruebas del ambiente configurado.
* HU5.- Mostrar las interfaces de las plataformas que entrega hadoop
* HU6.- Generar guía paso a paso de instalación de hadoop
* HU7.- Se requiere una plataforma en donde se puedan agregar, modificar y eliminar datos
* HU8.- Se requiere una interfaz en donde se puedan realizar el seguimiento de los datos
* HU9.- Generar la guía de instalación paso a paso de las herramientas para el manejo de datos

# Requerimientos de alto nivel

## Requerimientos Sistema:

* RS1 Permitir que múltiples usuarios pueden realizar diversas operaciones en una misma instancia de tiempo.
* RS2 Permitir el almacenamiento de datos
* RS3 Permitir modificación datos Ingresado.
* RS4 Permitir eliminar datos Ingresado.
* RS5 Permitir la realización de búsquedas de datos
* RS6 Permitir detener los servicios del ambiente hadoop cuando se estime conveniente
* RS7 Permitir al usuario el seguimiento de sus datos en una plataforma con interfaz

## Requerimientos Funcionales:

* RF1. El usuario debe ser capaz de leer la guía de instalación de la herramienta hadoop y sus servicios
* RF2. EL usuario debe ser capaz de iniciar la aplicación virtual box para llegar al ambiente hadoop
* RF3. La plataforma debe mostrar los servicios elevados para la utilización de la herramienta hadoop
* RF4. El usuario debe poder acceder a las plataformas de los servicios elevados y ver su interfaz de plataforma
* RF5. El usuario podrá realizar pruebas en el ambiente hadoop.
* RF6. El usuario debe ser capaz de ingresar, modificar y eliminar datos
* RF7. El usuario debe ser capaz de dar seguimiento a sus datos en la plataforma elevada por hadoop
* RF8. El usuario debe ser capaz de leer y entender la guía de instalación de las herramientas Apaches

## Requerimientos No Funcionales:

Los requerimientos no funcionales, se establecieron bajos las normas **ISO 9126**, el cual es un estándar para la evaluación de la calidad de software.

* RNF1. Funcionalidad: La plataforma debe cumplir con funciones:
  + Adecuación, exactitud, Interoperabilidad y Seguridad.
* RNF2 Fiabilidad: Un conjunto de atributos relacionados con la capacidad del software de mantener su nivel de prestación bajo condiciones establecidas durante un período establecido.
  + Madurez, Recuperabilidad, Tolerancia a fallos y Cumplimiento de Fiabilidad
* RNF3. Usabilidad: Un conjunto de atributos relacionados con el esfuerzo necesario para su uso
  + Aprendizaje, Comprensión, Operatividad y Atractividad
* RNF4. Eficiencia: Conjunto de atributos relacionados con la relación entre el nivel de desempeño del software y la cantidad de recursos necesitados bajo condiciones establecidas-
  + Comportamiento en el tiempo y Comportamiento de recursos
* RNF5. Portabilidad: Conjunto de atributos relacionados con la capacidad de un sistema de software para ser transferido y adaptado desde una plataforma a otra.
  + Capacidad de instalación y Capacidad de reemplazamiento
* RNF6. Calidad en uso: Conjunto de atributos relacionados con la aceptación por parte del usuario final y Seguridad.
  + Eficacia, Productividad, Seguridad y Satisfacción

## Delimitación

* Espacio. El espacio para la realización del proyecto será en la Universidad Andrés Bello de viña del mar, teniendo en cuenta que se busca implementar un ambiente de pruebas y desarrollo para Big Data.
* Tiempo. El planteamiento, desarrollo e implementación del proyecto será realizado durante 4 meses y medio.
* Alcance. El proyecto culmina con el diseño y montaje de una plataforma a base de las herramientas hadoop y los servicios de apache para el manejo de Big data.

Teniendo en cuenta que las herramientas y tecnologías serán analizadas y definidas como parte del proyecto, las herramientas que se van a implementar serán open source, permitiendo así que se puedan descargar y trabajar sin alguna restricción de licencia.

## El desarrollo del proyecto incluye:

* Un caso de estudio que permitirá realizar las pruebas necesarias del ambiente implementado.
* Guías paso a paso sobre la instalación, configuración e integración de cada una de las herramientas requeridas para la construcción del ambiente.
* Montaje de la herramienta hadoop con sus respectivos servicios Open Source.
* Montaje de las herramientas que trabajaran con hadoop “Herramientas Apache” con sus respectivos servicios

# Plan de proyecto

En este apartado se revisará el plan de proyecto, las metodologías utilizadas y la planificación estimada a fin de semestre

## Dirección de proyecto

Para dirigir este proyecto se utilizarán dos metodologías, la cuales trabajarán de forma separada para lograr la realización del proyecto.

**La metodología ágil Scrum,** se utilizará para la documentación y la forma de trabajo del proyecto, este ofrece un marco de buenas prácticas que favorece la adaptación los cambios, mitigación de riesgos, retroalimentación continua, etc.

Esta metodología trabaja con periodos de tiempos llamados Sprint, estos duran de 1 a 4 semanas

**FASES DEL SCRUM**

**INICIA**

**PLANEA Y ESTIMAR**

**IMPLEMENTAR**

**REVISION Y REROSPECTIVA**

**ANZAMIENTO**

Ilustración 4: Fases de SCRUM

SCRUM como cualquier metodología cuenta con diferentes fases, y cada fase con distintos procesos, las cual se especifican a continuación:

Iniciar es la primera fase de la metodología, donde se forman los equipos, se identifican y/o asignan los roles de los stakeholder (Personas afectadas por el proyecto, ya sea el equipo de desarrollo, socios, patrocinadores, etc.).

Los roles principales son:

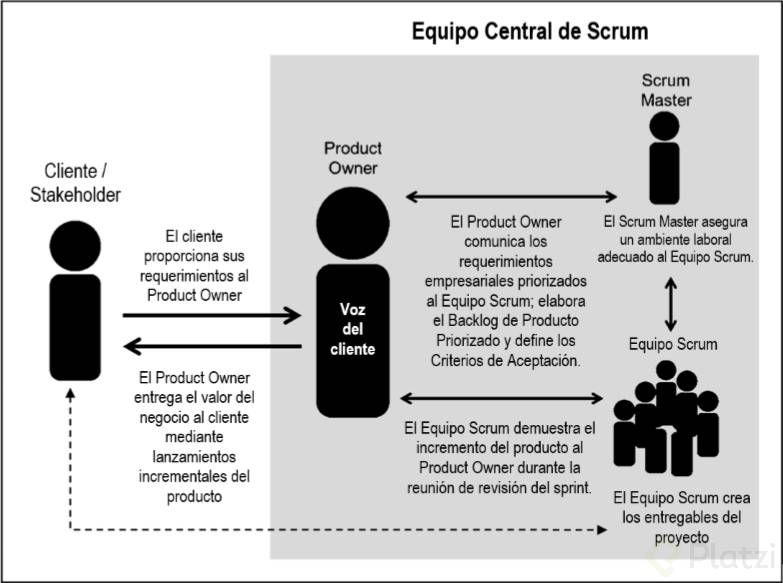


Ilustración 5: Equipo central de SCRUM

* **Scrum master:**Persona que lidera al equipo guiándolo para que cumpla las reglas y procesos de la metodología. Gestiona la reducción de impedimentos del proyecto y trabaja con el ProductOwner para maximizar el rol.
* **Product owner (PO):** Representante de lso accionistas y clientes que usan el software. Se focaliza en la parte de negocio y él es responsable del ROI del proyecto (entregar un valor superior al dinero invertido). Traslada la visión del proyecto al equipo, formaliza las prestaciones en historiasa incorporar en el Product Backlogy las reprioriza de forma regular.
* **Equipo de SCRUM:** Grupo de profesionales con los conocimientos técnicos necesarios y que desarrollan el proyecto de manera conjunta llevando a cabo las historias a las que se comprometen al inicio de cada sprint.

## Tabla de roles

|  |  |
| --- | --- |
| Rol | Responsabilidad |
| Cliente | **Sarita Gonzales** |
| Product Owner | **Sarita Gonzales** |
| Scrum Master | **David Córdova** |
| Equipo Scrum | **David Córdova** |

Tabla 3: Tabla de roles

## Posteriormente, se desarrollan varios documentos:

Documento de visión del proyecto: define el alcance y el objetivo de alto nivel de un programa, producto o proyecto. Una declaración clara del problema, una propuesta de solución y las características de alto nivel de un producto ayudan a establecer expectativas y reducir riesgos.

## Scrum

Lista de pendientes del producto (Product Backlog): es una lista de historias de usuario. Las historias de usuario, son pequeños documentos que identifican al usuario, su necesidad y finalidad, y estas fueron extraídas y pulidas de la épica, separando cada historia por requerimiento agrupándolas según necesidad.

El plan de lanzamiento y otros documentos opcionales.

En la fase planear y estimar, a partir de documentos de la fase anterior, se elaboran varios documentos los cuales servirán para estimar y planificar el proyecto. De la(s) épica(s) se elaboran las Historias de usuario (User stories), las cuales deben ser aprobadas por el Product Owner, para luego ser revisadas por el Scrum master junto con el equipo Scrum y estimar cuanto esfuerzo es necesario para desarrollar la funcionalidad descrita en cada historia y el valor que le da al negocio.

Para la estimación de esfuerzo se utiliza la técnica puntos de historia o puntos de función, la técnica puntos de historia consiste en estimar cuanto tiempo demora un solo miembro del equipo en desarrollar la historia en cuestión, un punto de historia es igual a una jornada de trabajo, si la historia es muy grande se utiliza Planning Poker. La técnica puntos de función, consiste en asignar puntos a la historia según la complejidad de los datos que se maneja y de los procesos que se realizan sobre ellos. Después de esto las historias de usuario son priorizadas y asignadas a sprint según el valor que aportan al negocio y el esfuerzo necesario. Cada historia de usuario es revisada y se identifican las tareas necesarias para cumplir con los requisitos de la historia. Como este es un proceso iterativo, se crea una lista de pendientes del sprint, es decir, historias de usuario o tareas que quedaron pendientes.

En la fase de implementación, al final de cada sprint, el equipo Scrum presenta un entregable. Un entregable es una o varias funcionalidades de software, el cual permite validar con el cliente la visión del proyecto y recibir una retroalimentación.

En esta fase, se llevan a cabo reuniones de equipo diarias (Daily Standup), las cuales permiten discutir el progreso o el enfoque del proyecto y por último llevar un mantenimiento del Product Back log, esta práctica no se utilizará, ya que el equipo de desarrollo cuenta con un solo miembro.

Respecto a la fase de revisión y retrospectiva, no se utilizará la práctica convocar Scrum de Scrum ya que es sólo un equipo de desarrollo, esto es realizar un seguimiento de su respectivo progreso, de los impedimentos del equipo y de las dependencias entre los equipos.

Reunión de revisión de sprint, en esta reunión el Scrum mastery el equipo Scrum se reúnen con el Product Owner, revisan el sprint y su entregable para retrospectiva del sprint y validan los criterios de aceptación y la aprobación del entregable.

Por último, el Scrum master y el equipo Scrum realizan una reunión de retrospectiva del sprint, donde discuten que hicieron bien, mal y que pueden mejorar. Esta práctica tampoco se utilizará ya que el Scrum master y el equipo Scrum son la misma persona.

## Product backlog del sprint

El producto backlog será realizado con los requerimientos funcionales del sistema.

Las prioridades son el orden en que se ejecutaran los requerimientos.

En el cuadro de historias de usuario se podrá visualizar con las historias de usuario están ligadas a los requerimientos.

Parar la priorización de los requerimientos se utilizará el método MoSCoW, esta es una técnica de priorización de requisitos basada en el hecho de que, aunque todos los requisitos se consideren importantes es fundamental destacar aquellos que permiten darle un mayor valor al sistema, lo que permite enfocar los trabajos de manera más eficiente.

A continuación, se podrá observar el significado de las priorizaciones

* **M – MUST** Requisitos que debe tener la solución. Son los requisitos mínimos que hacen usable la solución
* **S – SHOULD** Requisitos que debería incluir la solución. Requisitos importantes, pero no obligatoriamente necesarios
* **C – COULD** Requisitos que podría incluir la solución.
* **W – WON’T** Requisitos que no se van a hacer (por el momento), pero que en un futuro podrían incorporarse

|  |  |  |  |
| --- | --- | --- | --- |
| ID | Requerimiento | Prioridad | Historias de Usuario |
| RQF1 | El usuario debe ser capaz de leer la guía de instalación de la herramienta hadoop y sus servicios | C-COULD | H.U 1 H.U6 |
| RQF2 | EL usuario debe ser capaz de iniciar la aplicación virtual box para llegar al ambiente hadoop | S-SHOULD | H.U 1 |
| RQF3 | La plataforma debe mostrar los servicios elevados para la utilización de la herramienta hadoop | C-COULD | H.U 2 |
| RQF4 | El usuario debe poder acceder a las plataformas de los servicios elevados y ver su interfaz de plataforma | C-COULD | H.U 3 |
| RQF5 | El usuario podrá realizar pruebas en el ambiente hadoop | S-SHOULD | H. U 4 |
| RQF6 | El usuario debe ser capaz de agregar, modificar y eliminar datos | M-MUST | H.U.7 |
| RQF7 | El usuario debe capaz de dar seguimiento a sus datos en la plataforma elevada por hadoop | C-COULD | H.U8 |
| RQF8 | El usuario debe ser capaz de leer y entender la guía de instalación de las herramientas Apaches | M-MUST | H.U 9 |

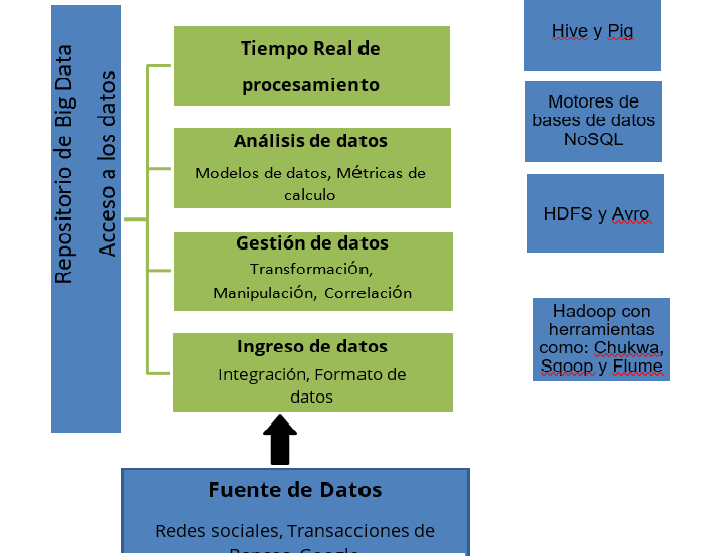
Tabla : Product Backlog

# Arquitectura de un ambiente de big data

## Arquitectura de Big Data

La gestión y procesamiento de Big Data es un problema abierto y vigente que puede ser manejado con el diseño de una arquitectura de 5 niveles, la cual está basada en el análisis de la información y en el proceso que realizan los datos para el desarrollo normal de las transacciones. A continuación, se pueden ver los niveles que contienen un ambiente Big Data y la forma en que se relacionan e interactúan entre ellos:

### Arquitectura de un ambiente de Big data

****

### 

Ilustración 6: Arquitectura big data

### Ingreso de datos

El Ingreso de datos es el procedimiento de obtener e importar información para su posterior uso o almacenamiento en una base de datos. Consiste en coleccionar datos de muchas fuentes con el objetivo de realizar un análisis basado en modelos de programación

### Gestión de datos

La administración de datos es el desarrollo y ejecución de arquitecturas, políticas, prácticas y procedimientos con el fin de gestionar las necesidades del ciclo de vida de información de una empresa de una manera eficaz. Es un enfoque para administrar el flujo de datos de un sistema a través de su ciclo de vida, desde su creación hasta el momento en que sean eliminados. La administración de Big data es la forma en que se organizan y gestionan grandes cantidades de datos, tanto de información estructurada como no estructurada para desarrollar estrategias con el fin de ayudar con los conjuntos de datos que crecen rápidamente, donde se ven involucrados terabytes y hasta peta bytes de información con variedad de tipos.

### Tiempo real de procesamiento

Es un proceso que automatiza e incorpora el flujo de datos en la toma de decisiones, este aprovecha el movimiento de los datos para acceder a la información estática y así lograr responder preguntas a través de análisis dinámicos. Los sistemas de procesamiento de flujo se han construido con un modelo centrado que funciona con datos estructurados tradicionales, así como en aplicaciones no estructuradas, como vídeo e imágenes.

El procesamiento de flujos es adecuado para aplicaciones que tiene tres características: calcular la intensidad (alta proporción de operaciones de E/S), permitir paralelismo de datos y por último la capacidad de aplicar los datos que se introducen de forma continua.

### Análisis de datos

Es el proceso de examinar grandes cantidades de datos para descubrir patrones ocultos, correlaciones desconocidas y otra información útil. Esta información puede proporcionar ventajas competitivas y resultar en beneficios para el negocio, como el marketing para generar mayores ingresos.

El objetivo principal del análisis de datos es ayudar a las empresas a tomar mejores decisiones de negocios al permitir a los científicos y otros usuarios de la información analizar grandes volúmenes de datos transaccionales, así como otras fuentes de datos que puedan haber quedado sin explotar por la inteligencia del negocio convencional.

### Bases de datos NoSQL

Con la aparición del término NoSQL en los 90s y su primer uso en el 2009 por Eric Vans, se pretende dar una solución a las problemáticas planteadas anteriormente, dando una posibilidad de abordar la forma de gestionar la información de una manera distinta a como se venía realizando.

Para dar una definición adecuada de las bases de datos NoSQL se pude tener en cuenta las siguientes características:

### Distribuido

Sistemas de bases de datos NoSQL son a menudo distribuidos donde varias máquinas cooperan en grupos para ofrecer a los clientes datos. Cada pieza de los datos se replica normalmente durante varias máquinas para la redundancia y alta disponibilidad.

### Escalabilidad horizontal

A menudo se pueden añadir nodos de forma dinámica, sin ningún tiempo de inactividad, lo que los efectos lineales de almacenamiento logran capacidades de procesamiento general.

### Construido para grandes volúmenes

Muchos sistemas NoSQL fueron construidos para ser capaz de almacenar y procesar enormes cantidades de datos de forma rápida.

### Modelos de datos no relacionales

Los modelos de datos varían, pero en general, no son relacional. Por lo general, permiten estructuras más complejas y no son tan rígida que el modelo relacional.

### No hay definiciones de esquema

La estructura de los datos generalmente no se define a través de esquemas explícitos que las bases de datos manejan. En su lugar, los clientes almacenan datos como deseen, sin tener que cumplir con algunas estructuras predefinidas.

# Hadoop y big data

Hadoop es un sistema de código abierto que se utiliza para almacenar, procesas y analizar grandes volúmenes de datos.

## Ventajas

* Aísla a los desarrolladores de todas las dificultades presentes en la programación paralela.
* Cuenta con un ecosistema que sirve de gran ayuda al usuario, ya que permite distribuir el fichero en nodos, que no son otra cosa que ordenadores con commodity-hardware.
* Es capaz de ejecutar procesos en paralelo en todo momento.
* Dispone de módulos de control para la monitorización de los datos.
* Presenta una opción que permite realizar consultas.

También potencia la aparición de distintos add- ons, que facilitan el trabajo, manipulación y seguimiento de toda la información que en él se almacena.

## Arquitectura hadoop

Dentro de la arquitectura de un ambiente Big Data se pueden utilizar diferentes herramientas, cada una de estas cumple un papel importante para la implementación. A continuación, se realiza una descripción de cada una de las tecnologías que están incluidas en la arquitectura.

## 

## Hadoop

Es un framework que permite el procesamiento distribuido de grandes conjuntos de datos a través de grupos de ordenadores que utilizan modelos de programación simple. Está diseñado para detectar y controlar los errores en la capa de aplicación.

Apache Hadoop tiene dos componentes centrales, el almacenamiento de archivos llamado Hadoop Distributed File System (HDFS), y la infraestructura de programación llamada MapReduce, como se muestra en la Figura.

## Arquitectura hadoop

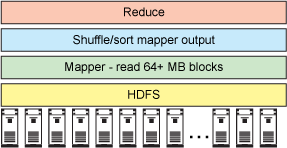


Ilustración : Arquitectura hadoop

## MapReduce

El modelo de programación MapReduce se basa en dos funciones llamadas Map y Reduce. La entrada a dicho modelo es un conjunto de pares clave/valor y la salida es otro conjunto de pares clave/valor.

### Función Map

A partir del conjunto de pares clave/valor de entrada se genera un conjunto de datos intermedios. La función Map asocia claves idénticas al mismo grupo de datos intermedios. Cada grupo de datos intermedios estará formado por una clave y un conjunto de valores, por lo tanto, estos datos intermedios van a ser a su vez la entrada de la función de Reduce.

### Función Reduce

La fase de Reduce se encargará de manipular y combinar los datos provenientes de la fase anterior para producir a su vez un resultado formado por otro conjunto de claves/valores.

Una fase intermedia es la denominada Shuffle la cual obtiene las tuplas del proceso Map y determina que nodo procesará estos datos dirigiendo la salida a una tarea Reduce en específico.

La siguiente figura da un ejemplo del flujo de datos en un proceso sencillo de MapReduce.

### Proceso MapReduce

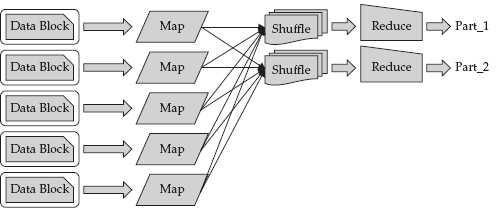


Ilustración : Arquitectura MapReduce

## HDFS

Es el sistema de ficheros distribuido utilizado por Hadoop. Por lo tanto, está especialmente diseñado para cumplir con las necesidades propias de Hadoop. Las dos ideas principales de HDFS es por un lado que sea un sistema de ficheros que facilite una alta escalabilidad tolerante a fallos. Por otro lado, Hadoop necesita que los problemas que se estén intentando solucionar involucren un gran número de datos. HDFS debe garantizar un alto rendimiento de datos para que Hadoop sea capaz de procesar.

A continuación, se puede observar la arquitectura que presenta HDFS con sus respectivos componentes:

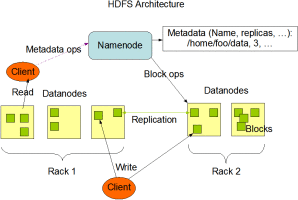


Ilustración : Arquitectura Hdfs

## Chukwa

Es un sistema de recopilación de datos de código abierto para el seguimiento de grandes sistemas distribuidos. Se construye en la parte superior del sistema de archivos distribuido Hadoop (HDFS) y Map/Reduce. Chukwa también incluye un conjunto de herramientas flexibles para la visualización, seguimiento y análisis de resultados de los datos recogidos.

La siguiente figura muestra la interacción de Chukwa con las otras tecnologías y la función que tiene dentro de los procesos de datos

### Diagrama de Chukwa

### 

Ilustración : Arquitectura chukwa

## Sqoop

Es una herramienta diseñada para transferir datos entre Hadoop y bases de datos relacionales. Sqoop importa los datos de un sistema de gestión de bases de datos relacionales (RDBMS) como MySQL u Oracle al sistema de archivos distribuido Hadoop (HDFS), donde transforma los datos y luego los exporta de nuevo a un RDBMS.

Sqoop automatiza la mayor parte de este proceso, basándose en la base de datos para describir el esquema de importación de los datos. Sqoop utiliza MapReduce para importar y exportar los datos, lo que proporciona el funcionamiento en paralelo, así como tolerancia a fallos.

## Flume

Es un servicio distribuido, confiable y disponible para recolectar, agregar y mover grandes cantidades de datos de registro eficientemente. Cuenta con una arquitectura simple y flexible basada en transmisión de flujos de datos. Es robusto y tolerante a fallos con los mecanismos de fiabilidad, conmutación por error y los mecanismos de recuperación. Se utiliza un modelo de datos extensible simple que permite una aplicación analítica en línea, el cual se puede ver en la siguiente figura.

### Flujo de datos

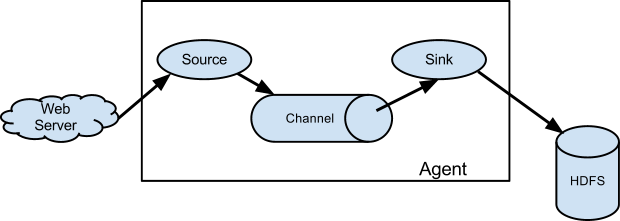


Ilustración : Flujo de datos flume

## Avro

Es un sistema de serialización de datos, contiene lo siguiente:

* Estructuras de datos.
* Formato de datos binario.
* Un archivo contenedor para almacenar datos persistentes.
* Llamada a procedimiento remoto (RPC).
* Fácil integración con lenguajes dinámicos. La generación de código no está obligado a leer y escribir archivos de datos ni a utilizar o implementar protocolos RPC.

Cuando los datos de Avro se almacenan en un archivo, su esquema se almacena con él, para que los archivos se puedan procesar posteriormente con cualquier programa. Si el programa de lectura de los datos espera un esquema diferente esto puede ser fácilmente resuelto, ya que ambos esquemas están presentes.

## Hive

Es la infraestructura de almacenamiento de datos construida sobre Apache Hadoop para proporcionar el resumen de datos, consultas ad-hoc y análisis de grandes conjuntos de datos. Proporciona un mecanismo para proyectar en la estructura de los datos en Hadoop y consultar los datos utilizando un lenguaje similar a SQL llamado HiveQL (HQL). Hive facilita la integración entre Hadoop y herramientas para la inteligencia de negocios y la visualización. Hive permite al usuario explorar y estructurar los datos, analizarlos y luego convertirla en conocimiento del negocio.

Estas son algunas de las características de Hive:

* Cientos de usuarios pueden consultar simultáneamente los datos utilizando un lenguaje familiar para los usuarios de SQL.
* Los tiempos de respuesta son típicamente mucho más rápido que otros tipos de consultas sobre el mismo tipo de conjuntos de datos.

Controladores JDBC y ODBC, aplicaciones para extraer datos. Hive permite a los usuarios leer los datos en formatos arbitrarios, usando SerDes y entrada / Formatos de salida.

## Pig

Es una plataforma para el análisis de grandes conjuntos de información que consiste en un lenguaje de alto nivel para la expresión de los programas de análisis de datos, junto con la infraestructura necesaria para la evaluación de estos programas. La propiedad más importante es que su estructura es susceptible de paralelismo, que a su vez les permite manejar grandes conjuntos de datos.

En la actualidad, la capa de infraestructura de Pig se compone de un compilador que produce secuencias de programas de Mapa-Reduce, para el que ya existen implementaciones paralelas a gran escala (por ejemplo, el Hadoop subproyectos).

### Pig posee las siguientes características:

#### Facilidad de programación

Es trivial para lograr la ejecución paralela de tareas de análisis simples. Las tareas complejas compuestas de múltiples transformaciones de datos relacionados entre sí, están codificados explícitamente como secuencias de flujo de datos, lo que hace que sean fáciles de escribir, entender y mantener.

#### Oportunidades de optimización

La forma en que se codifican las tareas permite que el sistema pueda optimizar su ejecución de forma automática, lo que permite al usuario centrarse en la semántica en lugar de la eficiencia.

#### Extensibilidad

Los usuarios pueden crear sus propias funciones para realizar el procesamiento de propósito especial

# Para la implementación de las herramientas

Para la implementación de alguna herramienta tecnológica en el proyecto, se utilizará la metodología tradicional iterativo-incremental.

## Metodología iterativo incremental

Ilustración : Metodología Iterativo-incremental

## ****Ciclo de vida iterativo****

Responde a la alta incertidumbre del proyecto realizando iteraciones, que no son más que una división del proyecto en fases cíclicas en las que el proyecto va avanzando progresivamente. A cada uno de estos ciclos se le denomina **iteración** y al inicio de cada una de ellas debe planificarse el trabajo a realizar en la misma.

Este ciclo de vida permite ir detallando el plan conforme avanza el proyecto y se va conociendo más sobre el mismo (disminuye la incertidumbre).

## ****Ciclo de vida incremental****

Es una particularización del anterior, mediante la cual cada ciclo que se realiza va obteniendo una porción de producto, servicio o resultado completa. A cada porción generada en una iteración se le denomina **incremento**.

Es decir, vamos produciendo porciones del resultado del proyecto que están acabadas al 100% e iteramos hasta tener todas las porciones, esto es, todo el resultado esperado.

El entregable se produce a través de una serie de iteraciones que sucesivamente añaden funcionalidad dentro de un marco de tiempo predeterminado. El entregable contiene la capacidad necesaria y suficiente para considerarse completo sólo después de la iteración final.

# Modelo de desarrollo

## Modelo Iterativo y acoplamiento al proyecto

Este modelo Busca reducir el riesgo que surge entre la necesidad del cliente y el producto final por procesos equívocos y confusos durante la etapa de recogida de requisitos

Este consiste en la iteración de varios ciclos de vida en cascada. Al final de cada iteración se le entrega al cliente una versión mejorada o con mayores funcionalidades del producto. El cliente es quien después de cada iteración evalúa el producto y lo corrige o propone mejoras. Estas iteraciones se repetirán hasta obtener un producto que satisfaga las necesidades del cliente.

Este tipo de modelo se utiliza en proyectos en los que los requisitos no están claros por parte del usuario, por lo que se hace la creación de distinta prototipos para presentarlos y conseguir la conformidad del cliente.

Este modelo servirá a este proyecto ya que la investigación de la tecnología big data y sus herramientas es algo totalmente nuevo y muy grande en el sentido de la implementación, los requerimientos de estas irán cambiando a medida que pase tiempo, ya que se irán incorporando nuevos requerimientos para ir completando y definiendo la investigación de este gran recurso de datos.

## Ventaja

Una de las principales ventajas de este modelo tradicional y que se acopla bastante bien al proyecto que estamos investigando, es que no hace falta que los requerimientos estén totalmente definidos al inicio del desarrollo, sino que se pueden ir refinando en cada una de las iteraciones en el tiempo.

## Desventajas

Las desventajas de este modelo, es que al no tener los requerimientos bien definidos del comienzo del proyecto la arquitectura del proyecto puede fallar y la implementación de la tecnología, ya que este proyecto dependerá netamente del mejor uso de herramientas tecnológicas que se acoplen de la mejor forma a las plataformas.

## Requerimientos de Hardware y software

* Sistema Operativo: CentOs Versión 6.10.
* Tipo de sistema operativo: 64 bits.
* JDK: Java 8.0 jdk
* JRE: Java
* Memoria RAM: 2 GB.
* Espacio libre en disco: 1 GB (Esto incluye únicamente la instalación, se puede requerir más dependiendo del tamaño de la data a procesar).