**Borrador de hito numero 2**

**Plataforma de soporte de proyectos big data**

**David Córdova Cordero**

# 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 |

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.

**Objetivo del proyecto**

**Planteamiento del problema**

**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.

**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**

* Desarrollar e implantar una plataforma en la nube para proyectos Big Data que permita dar soluciones a los problemas de las empresas, permitiendo el almacenamiento y uso de grandes tamaños de información, entregando versatilidad, seguridad y permitiendo que el usuario organice, analice y localice sus datos brindando una gran velocidad en la búsqueda y análisis de información a tiempo real para que este pueda tomar grandes decisiones en el menor tiempo posible.

Todo esto para incrementar al menos a un 80 % el uso de las plataformas de las plataformas que brinden estos servicios dando a conocer el cómo funcionan y la mejor manera de elegir la plataforma y herramientas adecuada para sus tipos de datos.

**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

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.

**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 del ambiente 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.

**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 como se acopla big data con las herramientas hadoop

**HU3** Mostrar funcionamiento y como se procesan los datos con las herramientas hadoop y sus herramientas adicionales.

**HU4** Realizar las pruebas del ambiente configurado.

**Haadop y big data**

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

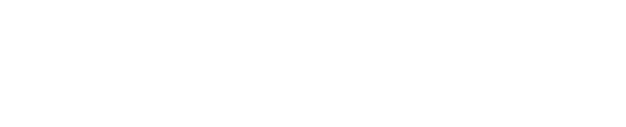
Ventajas

* 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 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



**Ingreso de datos**

Integración, Formato de datos

**Gestión de datos**

Transformación, Manipulación, Correlación

**Análisis de datos**

Modelos de datos, Métricas de calculo

**Tiempo Real de procesamiento**

Alertas y Triggers

**Fuente de Datos**

Redes sociales, Transacciones de Bancos, Google

HDFS y Avro

Motores de bases de datos NoSQL

Hive y Pig

Repositorio de Big Data

Acceso a los datos

Hadoop con herramientas como: Chukwa, Sqoop y Flume

**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.

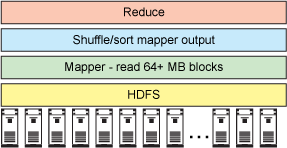
**Arquitectura hadoop**

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



**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.

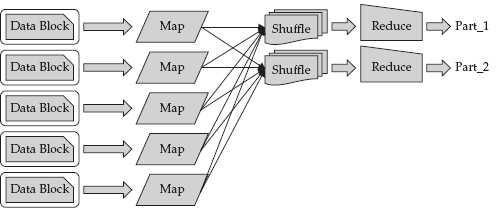
**Función Reduce**

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.

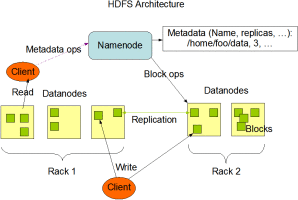
La siguiente figura da un ejemplo del flujo de datos en un proceso sencillo de MapReduce.

# Proceso MapReduce



**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:

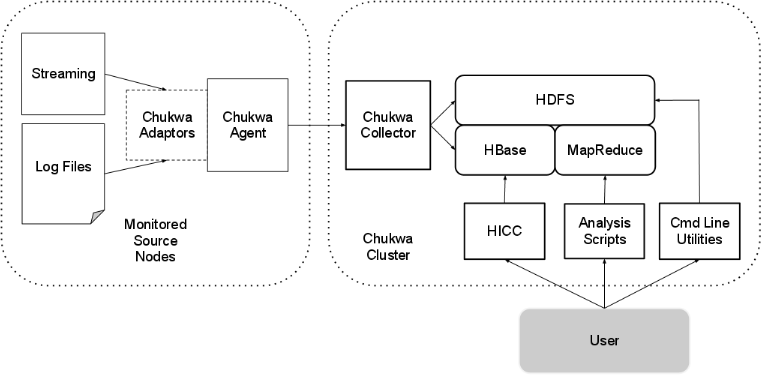


**Chukwa**

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



**Sqoop**

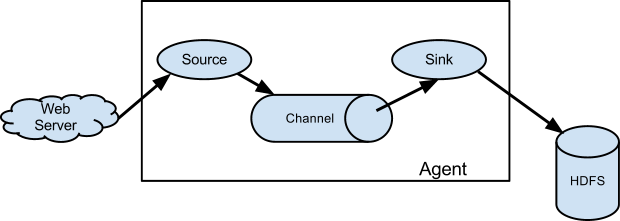
Es una herramienta diseñada para transferir datos entre Hadoop y bases de datos relacionales. Sqoop importa los datos de un sistema de gestión de bases de datos relacionales (RDBMS) como MySQL u Oracle al sistema de archivos distribuido Hadoop (HDFS), donde transforma los datos y luego los exporta de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



**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.

**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: Ubuntu 19.04.
* Tipo de sistema operativo: 64 bits.
* JDK: Java 8.0.
* 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).

# Pasos que seguir

* Instalar Hadoop con el fin de configurar un clúster y lograr incluir todas las herramientas en el mismo nodo.
* Instalar Hive (Framework para consultas SQL).
* Integrar Hive con Hadoop.
* Instalar Hbase dentro del cluster de Hadoop (motor NoSQL).
* Integrar Hbase con Hive.

