**UNIVERSIDAD ANDRES BELLO**

**FACULTAD DE INGENIERÍA**

**INGENIERÍA EN COMPUTACIÓN E INFORMÁTICA**



Plataformas de soporte de proyecto Big Data

Proyecto de titulo 1

**David Alejandro Córdova Cordero**

**VIÑA DEL MAR – CHILE**

**Abril 2019**

Contenidos

[1. Introducción 7](#_Toc5668268)

[2. Fundamentación del Problema 10](#_Toc5668269)

[2.1. Análisis de la situación actual. 10](#_Toc5668270)

[2.2. Análisis de la problemática. 11](#_Toc5668271)

[2.2.1. Técnica de los cinco por qué. 11](#_Toc5668272)

[2.2.2. Objetivo general. 12](#_Toc5668273)

[2.2.3. Objetivos específicos. 12](#_Toc5668274)

[2.2.4. Diagrama de alto nivel. 14](#_Toc5668275)

[2.3. Arquitectura de la solución. 15](#_Toc5668276)

[2.4. Solución ideal. 17](#_Toc5668277)

[2.5. Limitaciones. 17](#_Toc5668278)

[2.6. Restricciones. 17](#_Toc5668279)

[2.7. Alternativas de solución. 17](#_Toc5668280)

[2.7.1. Procedimientos manuales. 17](#_Toc5668281)

[2.7.2. Cambios en procedimientos actuales. 18](#_Toc5668282)

[2.7.3. Alternativas disponibles en el mercado. 18](#_Toc5668283)

[2.8. Solución propuesta. 19](#_Toc5668284)

[2.9. Situación Futura 19](#_Toc5668285)

[2.9.1. Factibilidad operativa 20](#_Toc5668286)

[2.9.2. Factibilidad técnica 20](#_Toc5668287)

[2.9.3. Factibilidad Económica 21](#_Toc5668288)

[2.10. Requerimientos de alto nivel. 21](#_Toc5668289)

[2.10.1. Requerimientos Sistema: 21](#_Toc5668290)

[2.10.2. Requerimientos Funcionales: 21](#_Toc5668291)

[2.10.3. Requerimientos No Funcionales: 22](#_Toc5668292)

[3. Planificación del proyecto. 24](#_Toc5668293)

[3.1. Metodología de desarrollo. 24](#_Toc5668294)

[3.1.1. Ventajas 26](#_Toc5668295)

[3.1.2. Inconvenientes 27](#_Toc5668296)

[3.1.3. Roles dentro del Proyecto 27](#_Toc5668297)

[3.2. Planificación del proyecto. 29](#_Toc5668298)

[3.2.1. Product Backlog 29](#_Toc5668299)

[3.2.2. Product Backlog 29](#_Toc5668300)

[3.2.3. Planificación del proyecto 29](#_Toc5668301)

[3.3. Plan general de Pruebas 29](#_Toc5668302)

[3.3.1. Responsables de las Pruebas 30](#_Toc5668303)

[3.3.2. Entorno de Pruebas 30](#_Toc5668304)

[Ambiente: 30](#_Toc5668305)

[Hardware: 30](#_Toc5668306)

[Software 30](#_Toc5668307)

[3.3.2.1. Entorno de desarrollo 31](#_Toc5668308)

[3.4. Gestión de la configuración. 31](#_Toc5668309)

[3.4.1. Etapas 31](#_Toc5668310)

[3.5. Gestión de cambios. 33](#_Toc5668311)

[3.6. Gestión de riesgos. 34](#_Toc5668312)

[Figura 3.4.1 Tabla valorización de riesgo 35](#_Toc5668313)

[3.7. Entorno de desarrollo. 36](#_Toc5668314)

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

Big data es un término que describe el gran volumen de datos ya sean estructurados como no estructurados que tomarían demasiado tiempo y sería muy costoso cargarlos una base de datos relacional para su análisis. De tal manera que, el concepto big data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales.

Es importante entender que las bases de datos convencionales son una parte importante y relevante para una solución analítica. De hecho, se vuelve mucho mas vital cuando se usa en conjunto con la plataforma de big data.

Para conocer las características de la palabra Big Data nos centraremos en explicar las 5V para así, comprender mejor el término de Big Data.

* + **Volumen**
* Es una de las características más destacadas y asociada al concepto de Big Data. La generación de datos en un día es superior a la de hace veinte años, todos esos datos desestructurados que se guardan tienen un inmenso potencial.  Por eso, para este gran volumen de información se requiere que se tenga una correcta orientación de las estrategias para poner filtrarlos y así ofrecer un ahorro de tiempo.
  + **Velocidad**
* Todo el proceso de tratamiento de datos pide agilidad para extraer el valor de negocio a toda la información que se tiene y que de esta manera no se pierda ninguna oportunidad. El tiempo de procesamiento de la información tiene que ser un factor fundamental para que el tratamiento aporte ventajas y marquen la diferencia.
  + **Variedad**
* Los datos que se recopilan pueden venir estructurados (son más fáciles de gestionar) o no estructurados (en forma de documentos, vídeos, mensajes de correo electrónico, redes sociales, etc.). Dependiendo de esta diferenciación, cada tipo de información se trata de manera distinta, mediante herramientas específicas, pero después la esencia del Big Data reside en combinar y configurar unos datos con otros.
  + **Veracidad**
* Existe mucha variedad en toda la información que se recaba; por ello es conveniente encontrar las herramientas necesarias para comprobar la veracidad de la información recibida.

Por ello, las empresas deben asegurarse de que los datos que están recopilando tengan validez, es decir, que sean los adecuados para los objetivos que se pretenden alcanzar con ellos.

* + **Valor**
* Toda la información recogida tiene que servir para aportar valor a las empresas, los gobiernos o la sociedad porque si no, no da lugar a almacenar ni administrar. En definitiva, la clave está en cómo obtener la mejor información, el mejor valor y conocimiento para sacar la mayor rentabilidad.

La naturaleza de la información hoy es diferente a la información en el pasado. Debido a la abundancia de sensores, micrófonos, cámaras, escáneres médicos, imágenes, etc. en nuestras vidas, los datos generados a partir de estos elementos serán dentro de poco el segmento más grande de toda la información disponible.  
El uso de Big Data ha ayudado a los investigadores a descubrir cosas que les podrían haber tomado años en descubrir por si mismos sin el uso de estas herramientas, debido a la velocidad del análisis, es posible que el analista de datos pueda cambiar sus ideas basándose en el resultado obtenido y retrabajar el procedimiento una y otra vez hasta encontrar el verdadero valor al que se está tratando de llegar.

Capítulo 2: Fundamentación del problema

# Fundamentación del Problema

# Análisis de la situación actual.

El big data se ha convertido en uno de los elementos recurrentes cuando se habla de cosas que debe hacer las marcas, sus responsables y de herramientas que pueden ayudar a mejoras los resultados de la empresa. Pero, a pesar de las grandes promesas de los datos y de lo mucho que se espera de ellos, lo cierto es que tiene una cara no muy positiva, ya que aún hay cosas que se siguen haciendo mal y que no permiten obtener los resultados que se espera. Las empresas siguen cometiendo los mismos errores, y uno de los más importantes es que los datos no están donde deben estar y no se pueden acceder a ellos ya que están desestructurados o mejor dicho dispersos lo que provoca la difícil tarea de localizarlos, analizarlos y controlarlos para su posterior uso en la toma de decisiones.

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 perdida 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.

La dificultad de localizar los datos no estructurados puede ser una cuestión de volumen. De hecho**, el 65% de las empresas recopilan tantos datos que no son capaces ni de categorizarlos ni analizarlos**. Con tal cantidad de datos, la dificultad de localizarlos todos puede resultar más compresible. Además, sin poder analizarlos, una empresa no sabrá ni su valor, ni qué medidas de seguridad hay que aplicar sobre ellos.

# 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?
   * 1. Por que manejan demasiado volumen de datos y no utilizan las herramientas y plataformas correctas acorde a los datos que maneja la empresa
2. ¿Por qué que es importante usar las herramientas y plataformas adecuadas para el buen uso de los datos?
   * 1. 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.
3. ¿Por qué utilizar plataformas que den soporte a proyecto big data?
   * 1. Por que 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.
4. ¿Por qué una plataforma sería la solución a la entrega de información tiempo real?
   * 1. 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.}
5. ¿Por qué big data suele ser un problema para las empresas que no utilizan herramientas para interactuar con estas?
   * 1. 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 parala 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.

**Problemática**

Los altos volumen de datos de big data y el mal uso de herramientas y plataformas de soportes provocan perdida de información y lentitud en el análisis de datos.

# 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.

1. Disminuir la perdida de datos de las empresas entregando una plataforma de soporte al gran volumen de datos

2. Disminuir los fallos en la toma de decisiones con mejores tiempos de respuesta en la entrega de información con una plataforma que albergue big data en la nube.

3. Localizar, procesar y analizar de formas más rápida y segura la información que se albergada en la plataforma

4. Disminuir las dificultades de entendimiento al momento de analizar información entregando una plataforma amigable para el cliente y fácil de entender con un lenguaje común.

5. Disminuir la probabilidad de robo de información en un 60 % entregando un estricto control de acceso y una fácil y rápida configuración de los niveles de permisos

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Objetivo específicos | Situación Actual | Objetivo especifico | Situación Esperada | Métricas | Criterio de Éxito |
| 1 | Empresas hacen uso de grandes volúmenes de datos sin plataformas que los manipulen provocando perdidas de estos | Disminuir la perdida de datos de las empresas entregando una plataforma de soporte al gran volumen de datos | Generar una plataforma con las tecnologías adecuadas para que maneje los volúmenes de datos de la mejor forma posible | Herramientas acordes los datos del cliente | >=5 |
| 2 | Los datos de las empresas son ilocalizables, provocando la demora en las respuestas y conllevando a malas decisiones. | Disminuir los fallos en la toma de decisiones con mejores tiempos de respuesta en la entrega de información con una plataforma que albergue big data en la nube. | Entregar información en rangos cortos de tiempo para poder ser ocupada cuando el administrador o encargado estime conveniente | Tiempo de espera – tiempo de respuesta | <=2 minutos |
| 3 | Las empresas no logran localizar los datos para poder procesarlos, esto provoca lentitud en generar información y poca seguridad ya que no se logra saber en donde se encuentran. | La plataforma debe localizar, procesar y analizar de formas más rápida y segura la información que se albergada en la plataforma | La información debe ser localizad de forma rápida para ser analizada y entregada de forma segura | Información localizada \*100 / información no localizada | >= 95% |
| 4 | Empresas ocupan plataformas inadecuadas poco amigables con el entorno del cliente provocando fallos y poco entendimiento en palabras técnicas. | Disminuir las dificultades de entendimiento al momento de analizar información entregando una plataforma amigable para el cliente y fácil de entender con un lenguaje común. | Plataforma con lenguaje y entorno amigable para el uso de cualquiera persona que no tenga conocimiento de lenguaje tecnológico | % de aprobación del entorno \*100 / desaprobación del entono | >= 80% |
| 5 | Información de las empresas se encuentran en lugares inadecuados con poca seguridad lo que provoca que cualquier persona tenga acceso a este. | Disminuir la probabilidad de robo de información en un 60 % entregando un estricto control de acceso y una fácil y rápida configuración de los niveles de permisos | Información albergada en los lugares adecuados con la seguridad y permisos a las personas correspondientes sin perder el uso amigable | N° de personas con permiso de ingreso\* 100 / N° de personas sin permiso de ingreso | >=90% |

Figura 2.3.2 Métrica

# 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.

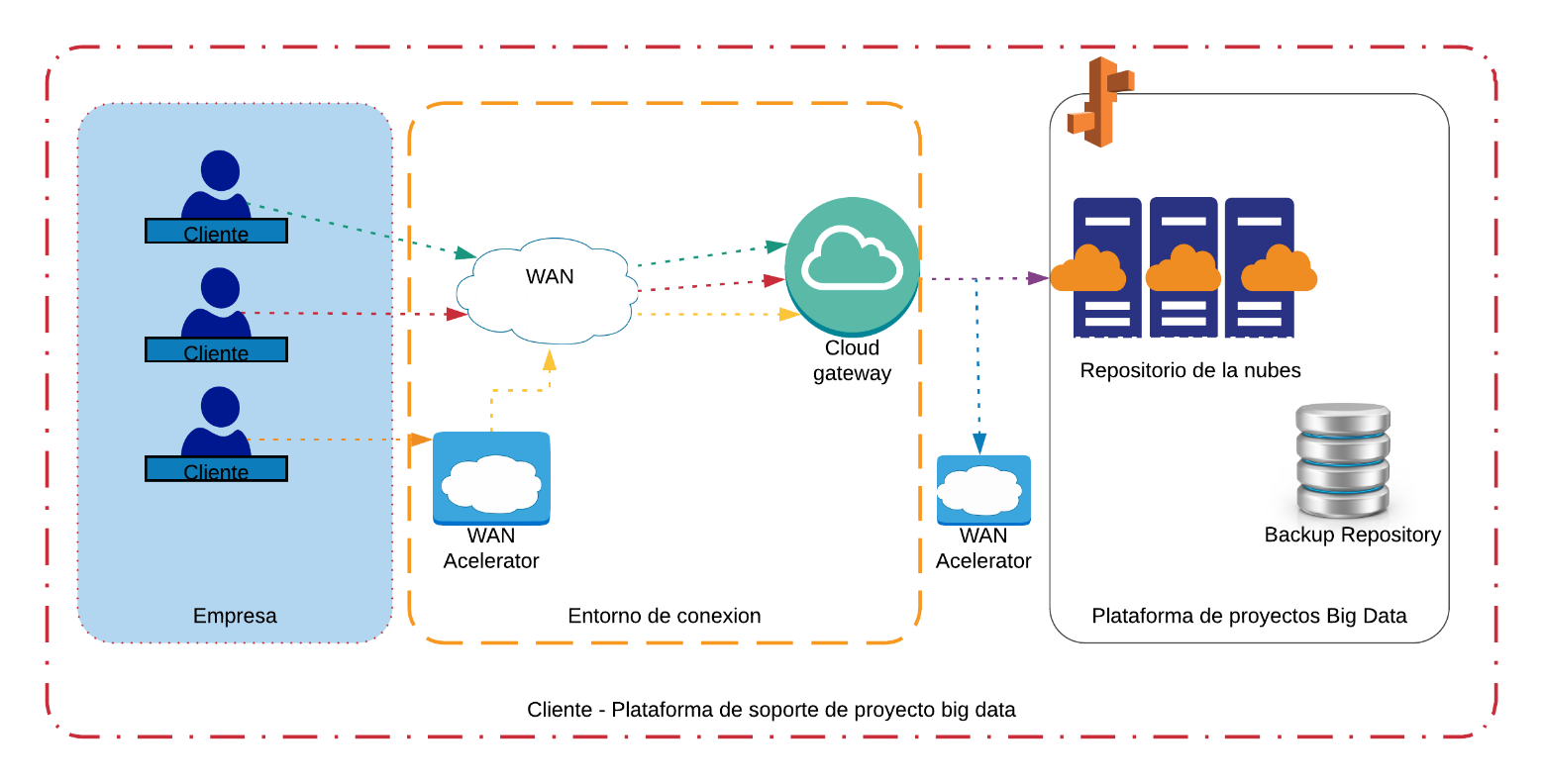


Figura 2.4.1 Diagrama de Alto Nivel

# Arquitectura de la 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.

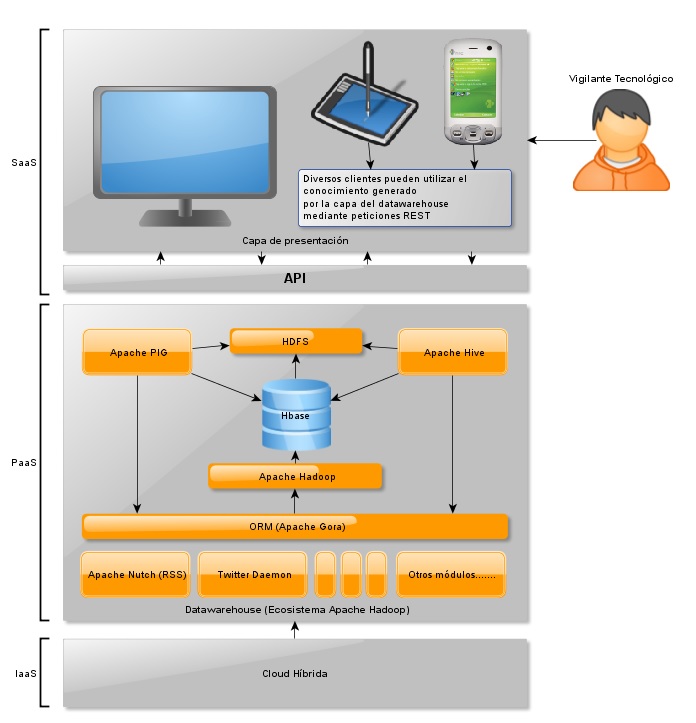
* Apache Pig: Plataforma de alto nivel para crear programas MapReduce utilizados en Hadoop
* Apache hive: Infraestructura de almacenamiento de datos construidas sobre hadoop para proporcionar agrupación, consulta, y análisis de datos.
* Apache mahout:Es un proyecto para producir implementaciones gratuitas 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

Figura 2.5.1 Diagrama tres capas

# Solución ideal.

El alcance de la solución propone que la creación de una nueva plataforma o mejora de las ya existente para entregar diseños más amigables y así dar mejores tiempos de respuesta para que los usuarios puedan entender de mejor forma su contenido y de esta forma tomar las mejores decisiones en conjunto a sus datos.

# 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 aplicación deberá mantener sesión de usuario.
* Los usuarios no tendrán acceso a internet
* El sistema deberá almacenar por un periodo de 3 meses las tarjas generadas.

# 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 ya existentes dejan de lado el entorno amigable y la seguridad de los datos del usuario.

Se debe modificar las plataformas dando un entorno mas amigable para la vista del cliente, incorporando lenguaje natural y no tecnicismos que solo entenderán las personas sabidas en el tema.

# 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 y muchos más cumplen por decirlo así casi las mismas funciones, las cuales son bastantes completas y /e funcionales para las empresas

# 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:

Hacer de estas plataformas su contenido un poco menos técnico y más amigable con el cliente, entregando un entorno que al cliente no le parezca difícil ni tedioso ocupar.

Entregar más seguridad y procesamiento en los datos, para que el cliente sienta que sus datos más importantes estén seguros. Esto se puede lograr mediante la investigación y el uso y/o combinación de herramientas que le den soporte a todos los tipos de datos que hoy en día existen ya sean datos estructurado como no estructurados.

Ofrecer multiplataformas 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 ya existente o que herramientas en conjunto puedo dar vida nuevas plataformas que entreguen al cliente todo lo que necesita para que pueda usar sus datos de la mejor forma.

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 debe solventar, 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.

# 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 La plataforma debe permitir generar reportes con datos
* RS7 La plataforma debe permitir al usuario ingresar aun cuando el sistema este en mantención
* RS8 La plataforma debe permitir crear gráficos con los datos
* RS8 Permitir al administrador ingresar nuevas personas con autenticación

# Requerimientos Funcionales:

* RF1. La plataforma debe solicitar usuario y contraseña de ingreso para acceder a la plataforma.
* RF2. La plataforma debe manejar perfiles de usuarios para limitar acceso entre los distintos tipos de usuarios.
* RF3. La plataforma debe usar validaciones para el correcto ingreso de la información antes de ser almacenada en la base de datos.
* RF4. La plataforma debe poder localizar, procesar y analizar los datos que el usuario quiera.
* RF5. El usuario debe poder acceder desde cualquier lugar con su usuario y contraseña
* RF6. El sistema debe de dar toda la seguridad a los datos guardando todos los registros de sus movimientos.

# Requerimientos No Funcionales:

* RNF1. Accesibilidad: La plataforma no debe ser difícil de utilizar por los distintos tipos usuarios.
* RNF2 Lenguaje: EL lenguaje utilizado en la plataforma debe ser lo más natural que se pueda
* RNF3. Disponibilidad: La plataforma debe estar operativo y funcional cada vez que un usuario lo disponga.
* RNF4. Seguridad: La plataforma debe respaldar los datos de manera confiable en un repositorio físico.
* RNF5. Rendimiento: El sistema debe ser rápido al localiza, analizar y procesar las órdenes que se le indiquen y obtener una rápida respuesta.

Capítulo 3: Planificación del proyecto

# Planificación del proyecto.

# Metodología de desarrollo.

Para el desarrollo de una plataforma de soporte para proyectos Big Data que brinde flexibilidad y escalabilidad, es necesario utilizar herramientas que permitan la culminación de trabajo de acuerdo con el tiempo planificado.

El presente proyecto utilizara el enfoque metodológico scrum, ya que este se enfoca en las prácticas de organización y gestión.

Al trabajar con la metodología Scrum sé que se libere un tipo producto al final de cada iteración. Las cuales están diseñadas para ser cortas y de duración fija. Este se enfoca en entregar una solución funcional cada poco tiempo.

Scrum es un proceso de gestión que reduce la complejidad en el desarrollo de productos para satisfacer las necesidades de los clientes. La gerencia y los equipos de Scrum trabajan juntos alrededor de requisitos y tecnologías para entregar productos funcionando de manera incremental usando el empirismo.

Los eventos de Scrum se utilizan para minimizar la necesidad de reuniones no definidas en Scrum y establecer una cadencia que permita al equipo fomentar la comunicación y colaboración reduciendo el tiempo en reuniones extensas además de reducir los procesos restrictivos y predictivos.

|  |
| --- |
| Resultado de imagen para scrum |
| Figura 3.1.1 Diagrama de Alto Nivel  Lo principal de Scrum es un Sprint, que es una iteración y puede durar un mes o menos, y se mantiene constante durante todo el desarrollo. Cada incremento del sprint brinda una funcionalidad utilizable al producto final. Cada sprint inicia inmediatamente después del anterior. En el Scrum todos los miembros del equipo tienen sus roles asociados y también existen los Bloques de Tiempo, Artefactos, y Reglas. Los bloques de tiempo crean orden y disciplina. Estos son: la Reunión de Planificación de la Entrega, la Reunión de Planificación del Sprint, el Sprint, el Scrum Diario, la Revisión del Sprint y la Retrospectiva del Sprint.  **Metodología Tradicional Prototipo**  En conjunto con scrum, la metodología que se adapta a este tipo de proyecto es la metodología tradicional de prototipo, la cual entregara versiones no terminadas del proyecto hasta llegar al proyecto terminado.  Las fases de esta son:  Planeación -Modelado-Elaboración del prototipo-Desarrollo-Entrega y retroalimentación-Comunicación con el cliente-Entrega del producto final.  **Por qué usar método prototipo**  La realización de una plataforma de desarrollo para proyecto Big Data, es un proyecto en el cual no se puede tener un resultado final de forma rápida, es un proyecto en el cual siempre surgirán nuevas actualizaciones que puedan mejorar el sistema, es por eso que se usar método prototipo es la mejor opción ya que este al permitir entrega de prototipo también permite en cada prototipo ir incorporando nuevas funcionalidades hasta llegara al producto final planeado. |

# Ventajas

**Modelo Scrum**

1. Gestión regular de las expectativas del cliente: El cliente establece sus expectativas indicando el valor que le aporta cada requisito del proyecto y cuando espera que esté completado.
2. Productividad y calidad: De manera regular el equipo va mejorando y simplificando su forma de trabajar.
3. Resultados anticipados: El cliente puede empezar a utilizar los resultados más importantes del proyecto antes de que esté finalizado por completo.
4. Mitigación de riesgos: Desde la primera iteración el equipo tiene que gestionar los problemas que pueden aparecer en una entrega del proyecto. Al hacer patentes estos riesgos, es posible iniciar su mitigación de manera anticipada.
5. Flexibilidad y adaptación: De manera regular el cliente redirige el proyecto en función de sus nuevas prioridades, de los cambios en el mercado, de los requisitos completados que le permiten entender mejor el producto, de la velocidad real de desarrollo, etc.

**Ventajas de Modelo tradición Prototipo**

1. Permite el desarrollo de un sistema a partir de requisitos pocos claros o cambiantes.
2. Son más fáciles de abordar con los usuarios finales.
3. Permite a todos los usuarios un mayor conocimiento del sistema con una menor de aprendizaje.
4. Permite a todos los involucrados entender bien y mejor el problema antes de implementación final.

# Inconvenientes

El inconveniente de este proyecto es que al ser un proyecto de investigación e implementación aún no se tienen clientes específicos que den a saber las carencias reales en sus respectivas empresas.

# Resultado de imagen para roles de scrumRoles dentro del Proyecto

**Roles y responsabilidades**

|  |  |
| --- | --- |
| Rol | Responsabilidad |
| Cliente | Sarita González |
| Product Owner | Sarita González |
| Scrum Máster | David Córdova |
| Equipo Scrum | David Córdova |

# 

# Planificación del proyecto.

.

# Planificación del proyecto

A continuación, se presenta la planificación del proyecto desde el punto de vista de la duración de los Sprint.

Para llevar a cabo el funcionamiento del proyecto se utilizó la herramienta Microsoft Project 2016, esta permite visualizar el sprint con sus respectivas tareas que serán desarrolladas, Mostrando como se distribuyen atreves del tiempo.

Como aún no se tiene estimado cuanto durará el proyecto, se realizará la planeación de la primera semana la cual dio inicio el 15 de marzo dando termino el 10 de abril. A continuación, se mostrara una imagen con los detalles del primer sprint.

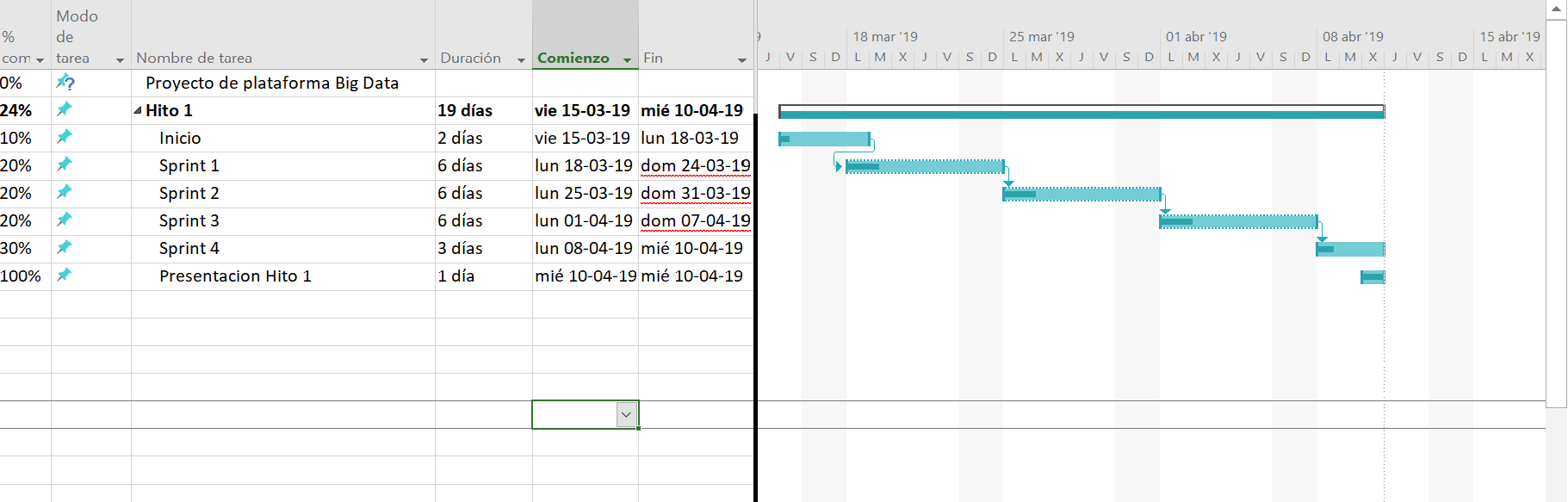


Figura 3.1.4.2 Planificación Carta Gantt

**Plan de entregable**

Cada sprint finalizará con un entregable que será liberado en el ambiente de prueba definido.

A continuación, a tabla de entregables del primer hito donde se definen que versiones y documentos serán liberados al final de cada sprint, y en qué ambiente estarán disponibles

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SPRINT | FECHA | VERSION | DOCUMENTOS | AMBIENTE | RESPONSABLE |
| 1 | 24/03/2019 | V1.0 | Prueba de herramientas BD | Pruebas | David Córdova |
| 2 | 31/03/2019 | V1.1 | Prueba de herramientas BD | Pruebas | David Córdova |
| 3 | 07/04/2019 | V1.2 | Prueba de herramientas BD | Pruebas | David Córdova |
| 4 | 10/04/2019 | V1.3 | Prueba de herramientas BD | Entregable | David Córdova |

En el primer sprint se deben tener pruebas de aceptación, las cuales son definidas por el Product Owner, al final de cada sprint el producto owner utiliza estos criterios para verificar los entregables completados, y para aceptar o rechazar los entregables.

Para llevar el plan de pruebas se generará un reporte de pruebas con el cual podremos ver:

* + Fechas ejecución
  + Tipos de pruebas
  + Versión de sistema probado
  + Estadísticas

# Plan general de Pruebas

# Responsables de las Pruebas

Las pruebas serán realizadas por el Product Owner y personas que el estime convenientes: Profesora Sarita Gonzales Catalán quien validara las funcionalidades.

# Entorno de Pruebas

# Ambiente:

Las pruebas de se desarrollará de la siguiente forma:

1. Alojamiento externo donde se realizarán las pruebas por parte de nuestro Product Owner.

# Hardware:

Servidor de propiedad de la empresa, en donde se aloja la base de datos SQL Server 2016 y el ambiente web en servidor de aplicaciones

# Software

* Bases de datos Mysql
* Aplicación de la nube sourceforge
* Framework Apache hadoop,
* Repositorio TortoiseSVN.

**Equipo.**

* Ordenador con acceso a internet y discos duros.

# Entorno de desarrollo

Se desarrolla en la dependencia de la Universidad Andrés Bello facultad de Ingeniería departamento de desarrollo, donde se proporcionarán los equipos y servicios pertinentes para el funcionamiento del proyecto, al igual que los ambientes respectivos bajo normas de seguridad.

# Gestión de la configuración.

Para la gestión de la configuración de este proyecto será divida en dos partes, gestión de líneas base y control de versiones.

**Gestión de línea Base**

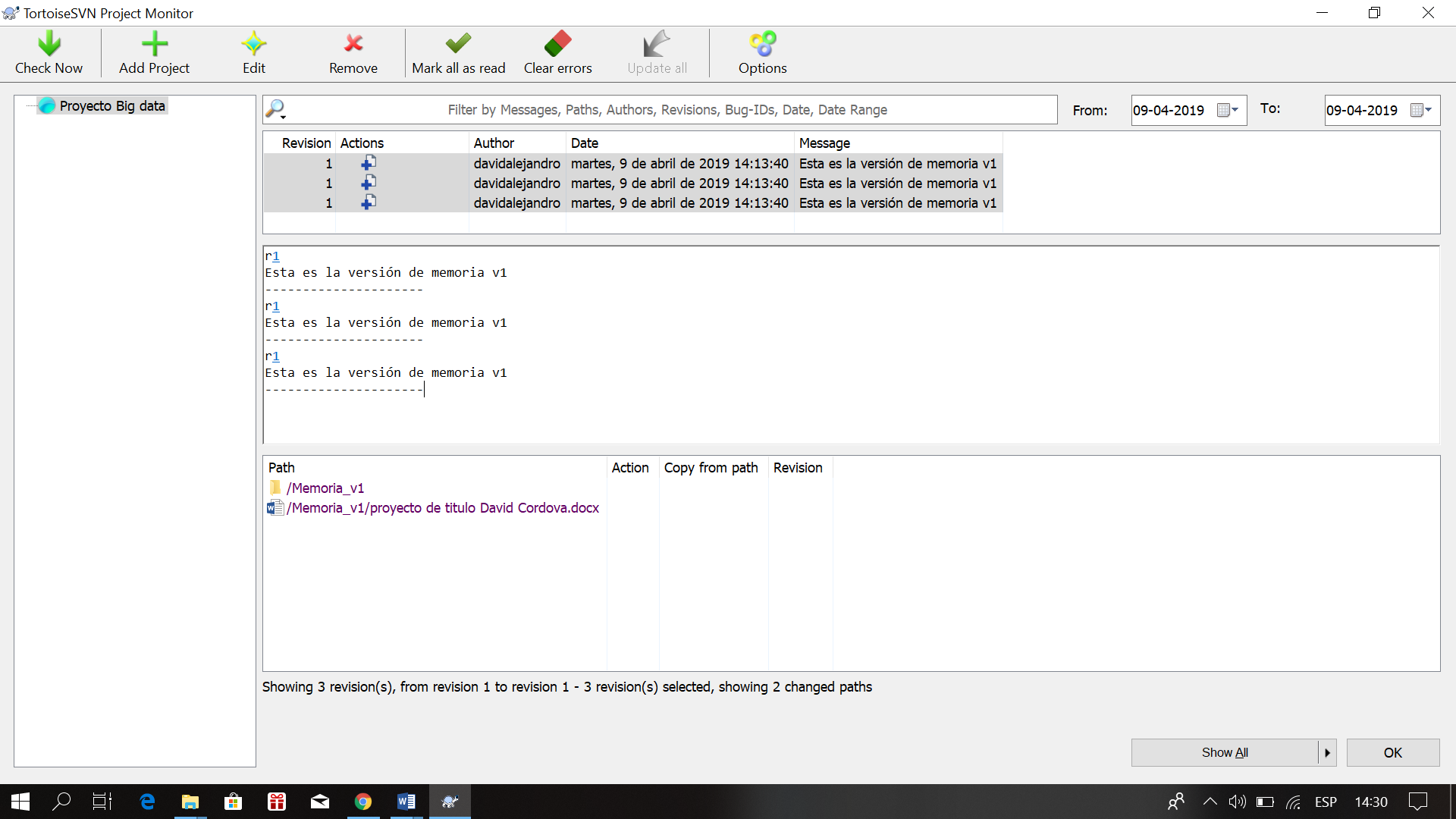
* Las líneas base para el desarrollo del proyecto serán establecida de la siguiente forma:
* Cada vez que se finalice una etapa será liberada una línea base del proyecto

Cada vez que se termine un sprint y este sea validado por el scrum Master se liberara una línea base del proyecto

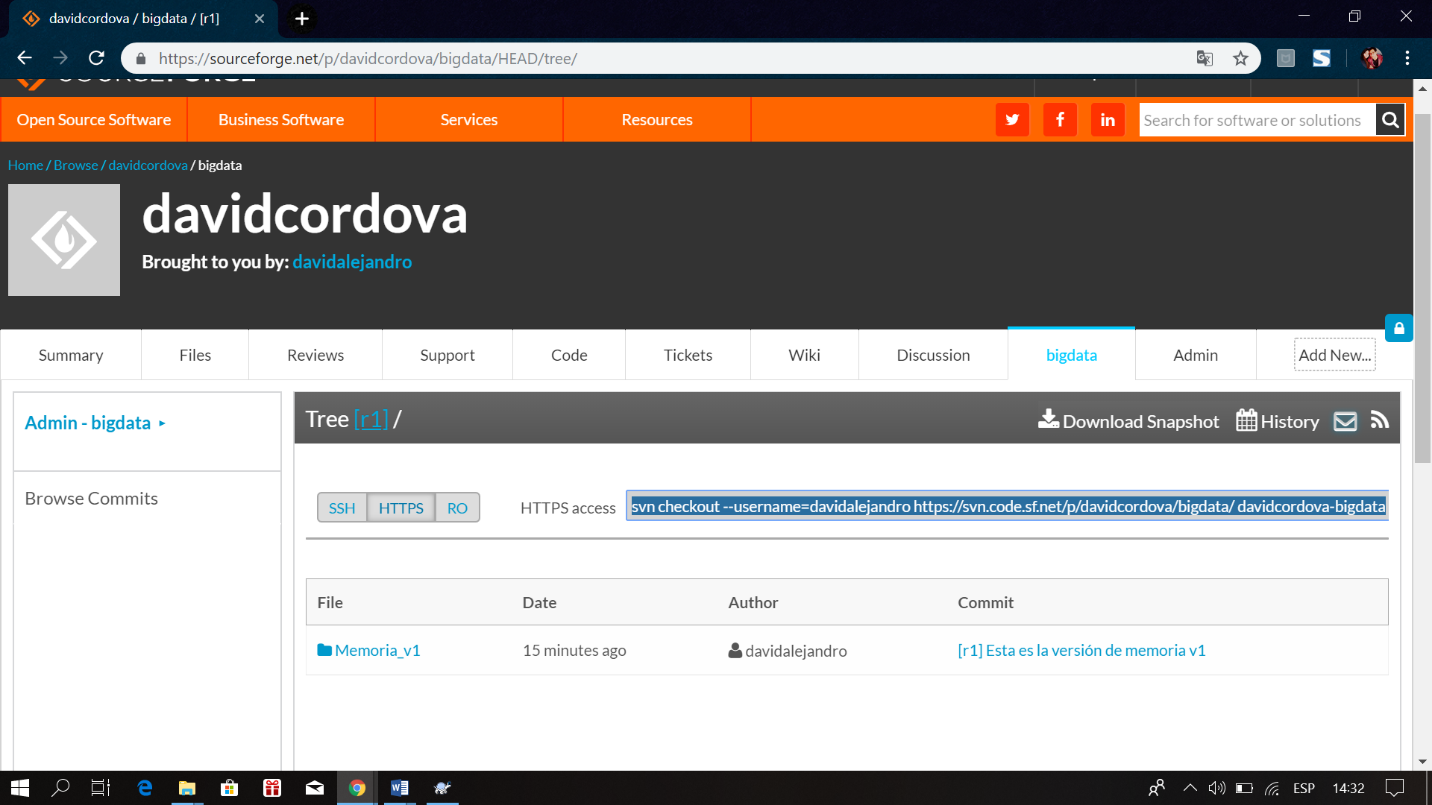
Presentación a evaluaciones (Hitos)

Durante el desarrollo del proyecto, se realizarán presentaciones ante una comisión para evaluar el avance del proyecto de título, para este propósito será liberada una línea base del proyecto.

Cada línea base del proyecto será documentada en el repositorio de la aplicación TortoiseSVN la cual conectará directamente con el repositorio en la nube sourceforge.net. Véase a continuación.



Conexión con Nube sourceforge.net/ la cual estar en conexión TortoiseSVN, en este apartado se puede ver reflejado que se encuentra guardara la versión 1 de memoria de título 1. Véase a continuación.



**Control de versiones**

Todos los documentos estarán controlados por versiones de desarrollo.

* Los documentos creados por primera vez serán asignados como v1.0(Versión 1.0)
* La versión será terminada una vez se tenga la aprobación del producto owner.
* Cuando un objeto sea modificado se generará una nueva versión. Ejemplo (v1.1).
* En caso de ser agregada un segundo hito al proyecto, este creara una nueva versión. Ejemplo (v2.0)

# Etapas

A continuación, se describen las etapas del proceso de control de cambios en términos de sus objetivos, criterios de entrada/salida y de sus actividades. Además, se enuncian los participantes de cada etapa y sus responsabilidades durante ella.

**Proponer cambios**.

**Objetivo**

Esta actividad tiene como objetivo el seguimiento del ciclo de vida de un cambio, que comprende: la solicitud del cambio, la evaluación, la aprobación o rechazo del mismo y la implementación.

# Gestión de cambios.

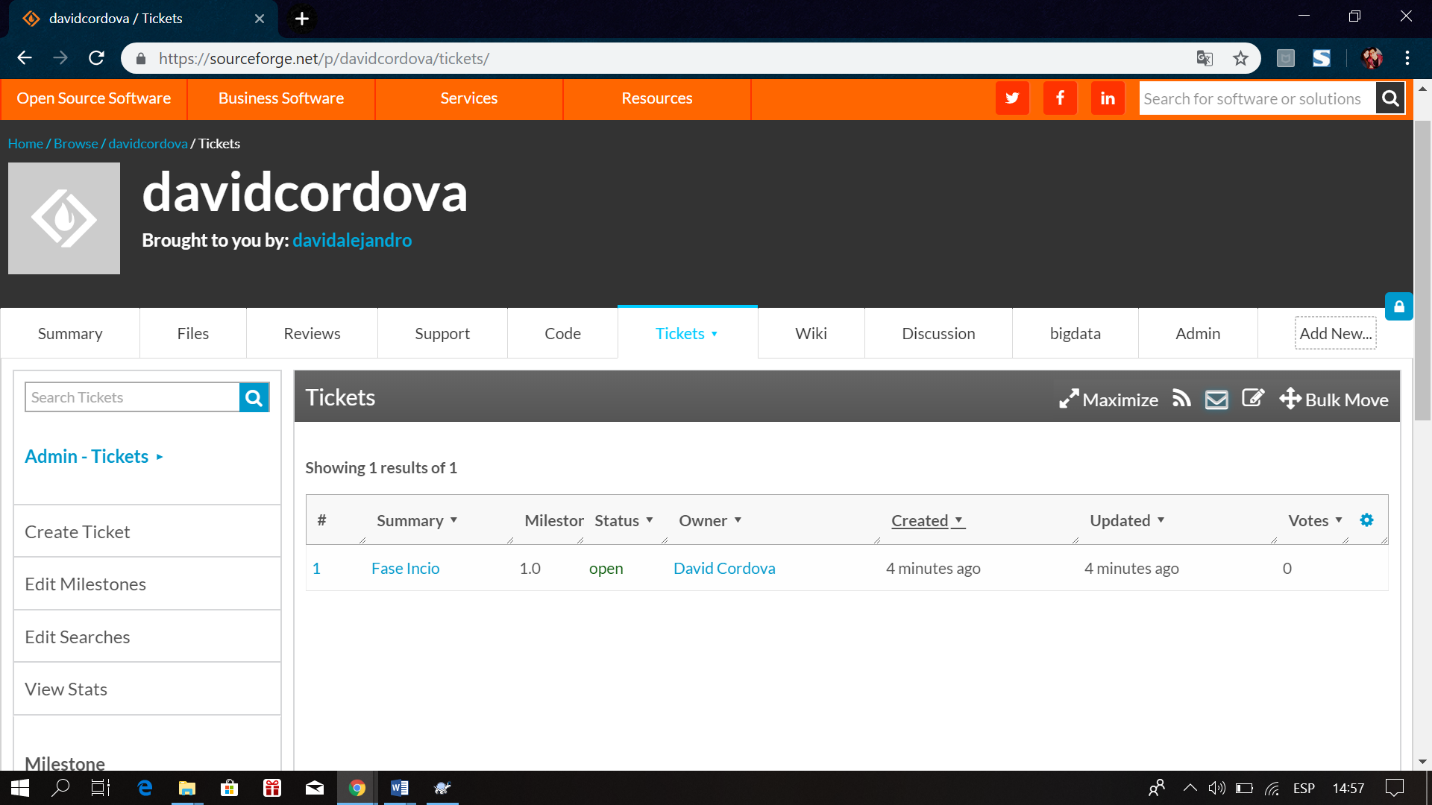
El proceso de la gestión de cambios será realizado a través de la nube en el siguiente enlace <https://sourceforge.net/p/davidcordova/cambios/>, donde se abrirán tickets cada vez que surja un problema entre el equipo de desarrollo.

Cada ticket pertenece a un "Milestone", este define un periodo de tiempo, por lo general será utilizado para los Sprint a excepción del Milestone 1. El cual es utilizado para la fase de Inicio.

El Milestone 2 corresponde al Sprint 1, el Milestone 3 corresponde al sprint 2 y así sucesivamente.

El ciclo de vida de un ticket tiene los siguientes estados:

* Nuevo ticket: El ticket ha sido creado especificando todas sus características principales (las cuales se describen más adelante en el documento).
* Asignado: El ticket ha sido asignado a un desarrollador (En este caso es cuando el equipo cuenta con más de un desarrollador los cuales se dividen las tareas).
* Cerrado El ticket ha sido cerrado, es decir terminado, resuelto.
* Re-abierto: El ticket ha sido re-abierto por diferentes circunstancias.



Se puede observar que cada modificación y/o cambio que se quiera procesar será ingresado directamente en la nube y con el repositorio TortoiseSVN.

Figura 3.3Control del Cambio

# Gestión de riesgos.

Un riesgo no necesariamente tiene un impacto negativo, los riesgos se clasifican como oportunidades cuando su impacto es positivo y amenazas cuando el mismo es negativo. Este plan se concentra en las amenazas y se han clasificado en tres categorías: Riesgos Administrativos (RA), son aquellas situaciones cuyo disparador depende de circunstancias propiamente ligadas al área administrativa. Riesgos Técnicos (RT), son aquellas situaciones cuyo disparador depende de circunstancias propiamente ligadas al área técnica. Riegos Externos (RE), son aquellas actividades donde el disparador del riesgo no depende de los departamentos o áreas de la organización, sino de un ente externo como un proveedor de bienes o servicios. Para lograr priorizar los riesgos identificados se utiliza una tabla de prioridad por impacto:

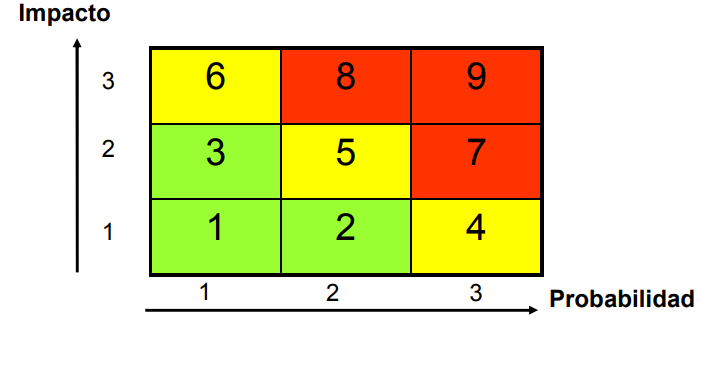
|  |  |  |  |
| --- | --- | --- | --- |
| ID | Riesgos | Mitigación | contingencia |
| RT1 | Desconocimientos de herramientas o tecnología a utilizar | Estudiar las diferentes alternativas. |  |
| RT2 | Mala elección a una metodología acorde al proyecto | Tabla comparativa de las distintas metodologías a usar. | Cambiar de metodología, adecuando las actividades ya realizada |
| RT3 | Problemas en el modelo de base de datos. | Análisis del modelo semanalmente. | Cambio del modelo mediante las herramientas |
| RT4 | Problemas en el perfil de usuario | Estudiar alternativas diferentes | Modificar código |
| RT5 | Problemas en el código | Realizar pruebas unitarias semanales | Realizar cambio de línea de comando errónea |
| RT6 | Problemas con inserción de información a la plataforma | Analizar y estudiar las alternativas de inserción de información a la plataforma | Modificar la forma de inserción de datos |
| RT7 | Problemas en la búsqueda de información | Analizar Rastreador de búsqueda de información | Modificar el rastreador de búsqueda de información |
| RA1 | Problemas de código | Realización de pruebas unitarias de forma semanal | Realizar los respectivos cambios a la línea de código errónea |
| RA2 | Comunicación con el cliente | Mostrar avances diarios | Comunicarse mediante correo, redes sociales u/o teléfono |
| RA3 | Cambios provenientes del cliente | Analizar posibles soluciones | Utilizar herramientas que el cliente solicite |
| RA4 | Atraso en actividades pauteadas | Trabajar de forma diaria en el proyecto | Aumentar hora diarias de trabajo |
| RE1 | Problemas con los servidores | Realizar mantenimiento a los servidores | Modificar servidores y servicios |
| RE2 | Problemas con la conexión a internet | Comunicarse con proveedor de conexión para solución | Cambiar de proveedor de conexión a internet |
| RE3 | Robo del material del proyecto | Buscar respaldo de la información del proyecto | Aumentar hora de trabajo sobre entregables pasados |
| RT8 | Investigación del proyecto mal realizada | Verificar fuente de donde se extrajo la información | Realizar investigación en más paginas para corroborar información |
| RT9 | Mala elección de repositorio de datos | Investigar la mejor opción de repositorio acorde al proyecto |  |

Figura 3.4.2 Tabla de riesgo

**Matriz de Probabilidad**

La matriz de probabilidad – impacto es una herramienta de análisis cualitativo de riesgos que nos permite establecer prioridades en cuanto a los posibles riesgos de un proyecto en función tanto de la probabilidad de que ocurran como de las repercusiones que podrían tener sobre nuestro proyecto en caso de que ocurrieran.

Como se puede ver en la siguiente ilustración, la matriz se compone de dos ejes: un eje vertical en donde se establecen los valores de probabilidad y un eje horizontal en donde se establecen los valores del impacto del riesgo sobre los objetivos de nuestro proyecto (en donde 0 implica que ese riesgo no repercutiría en los objetivos y 1 que dificultaría en gran medida el cumplimiento de los mismos). Los valores obtenidos en las diferentes celdas de la matriz son el resultado de multiplicar la probabilidad de ocurrencia por el impacto del riesgo, indicando los valores más altos (máximo 1) los riesgos más críticos del proyecto y los más bajos los menos relevantes.



# Figura 3.4.1 Tabla valorización de riesgo

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| N° | Riesgos | Probabilidad | Impacto | Nivel |
| 1 | Desconocimiento total de la plataforma en cual se desarrolla el primer proyecto | 3 | 3 | 9 |
| 2 | Dificultad en el uso de las herramientas que dan conexión las plataformas | 2 | 2 | 6 |
| 3 | Cambio en los objetivos generales y específicos | 1 | 3 | 3 |
| 4 | Nuevos requerimientos de alto nivel | 2 | 3 | 5 |
| 5 | Elección de mala metodología para el proyecto | 1 | 3 | 2 |

# Entorno de desarrollo.

Se realizarán pruebas de levantamientos de un repositorio TortoiseSVN conectado directamente a la nube de sourceforge, en la cual trabajaremos con conexiones MySQL.