**UNIVERSIDAD ANDRES BELLO**

**FACULTAD DE INGENIERÍA**

**INGENIERÍA EN COMPUTACIÓN E INFORMÁTICA**



Proyecto Big Data “Ecosistema Apache Hadoop”

Seminario de Título 2

**David Córdova Cordero**

**VIÑA DEL MAR – CHILE**

**Agosto 2019**

**Anexo**

**Introducción**

En el siguiente documento, se entregará la información necesaria y concisa de la memoria de proyecto de título 2, en este se detallará como se pulirán ciertas tareas para dar cumplimientos a las historias de usuario que no fueron realizadas en proyecto de título 1.

Para entregar un ecosistema acorde a lo planteado en el objetivo general del proyecto, se hará instalación de herramientas open source de apache que se complementen con las ya instaladas en PT1, están entregaran funcionalidad que el cliente planteo como historias de usuario y estarán sujetas a los entandares de calidad que se plantearan en el documento.

**Plan de proyecto seminario de título 2**

El plan de proyecto que se podrá ver a continuación se basara en las historias de usuario y las tareas que no se desarrollaron en el anterior seminario de título, este plan de proyecto se desarrolla para cumplir con los tiempos estimados por el cliente. El plan de proyecto se realizará con Microsoft Project 2016 en el cual se podrá ver a continuación las fechas estipuladas para el desarrollo de cada tarea para dar cumplimiento de a las historias de usuario planteadas. Se hará uso de métricas de gestión y desarrollo para poder mostrar cómo se comportó el manejo de las historias, para esto se hará uso de los gráficos Burndown y velocidad puntos de historias de usuario.

La planificación del seminario de título estará compuesta por 3 Sprint de 4 semanas, 2 sprint para la configuración, codificación y termino de memoria del proyecto y 1 sprint para el planteo y propuestas de estándares de calidad para el proyecto, en la siguiente imagen de carta Gantt se podrá observar la planificación del seminario.

Para el plan de prueba se generará un reporte de pruebas en el cual se podrá observar los siguientes:

* Fecha de ejecución de la prueba
* Tipo de pruebas realizada
* Versión del sistema

A continuación, se podrá observar la tabla del reporte.

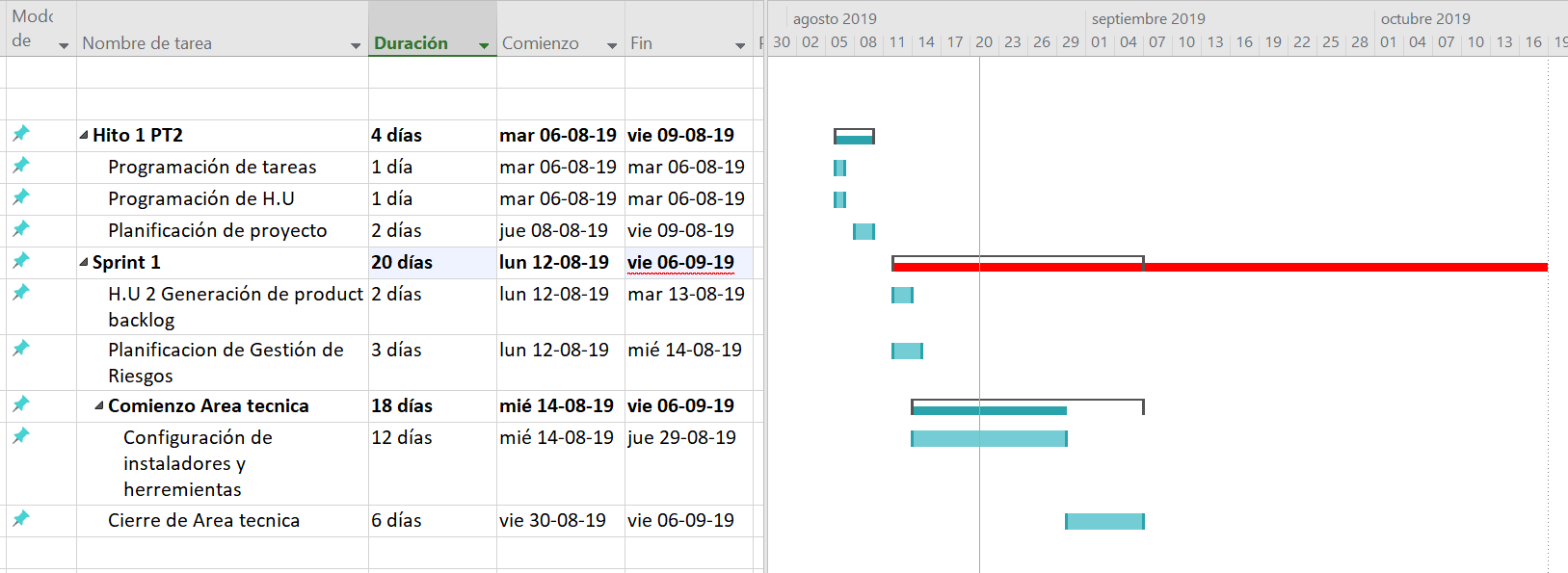
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| ID PRUEBA | H.U | NOMBRE | DESCRIPCION | ENTRADA | RESULTADO | RESULTADO | Estado |
| P.0.1 | H.U.2 | Prueba de funcionamiento de nuevo ecosistema HUE | Pruebas de códigos en archivos XML.  Pruebas de la plataforma de HUE | -Revisión de códigos funcionales  -Revisión de levantamiento de la nueva interfaz |  | Levantamiento de la interfaz de apache HUE | REVISIÓN |
| P.0.2 | H.U.2 | Pruebas de manejo de datos en HUE | Incorporar datos al ecosistema de HUE para realizar procesamiento | Aplicar estándares de calidad a los datos procesados |  | Datos procesados exitoso | REVISION |

El estado de cada prueba realizada a las historias de usuario puede ser aprobado o rechazado.

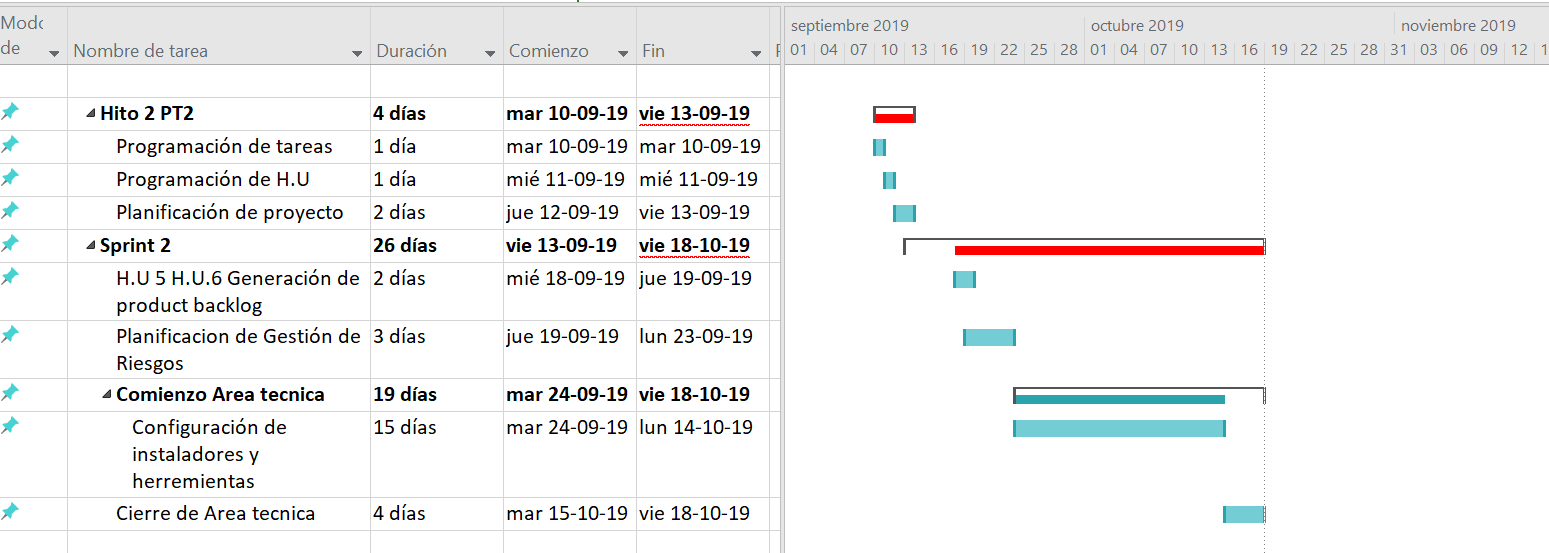
El estado aprobado da a conocer que la prueba se ha aprobado sin ningún inconveniente, por otra parte, el rechazado da conocer que la prueba tubo fallas las cuales no se acercan al resultado esperado.

**Planificación de Hito 1, Sprint 1:**

A continuación, se mostrará cómo se trabaja en el hito 1 y como se desarrollará la historia de usuario 2, los tiempos que serán destinados para cumplir con los plazos que el cliente estipulo

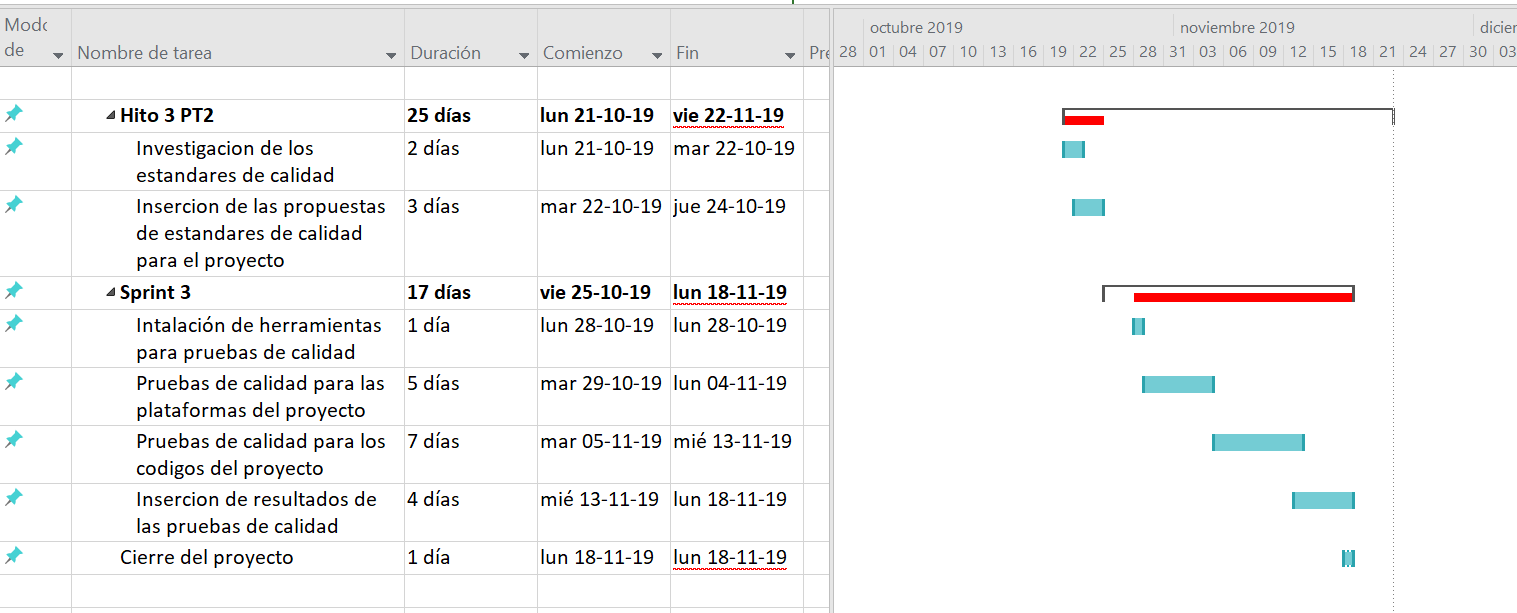


**Planificación de hito 2, Sprint 2:**

A continuación, se mostrará cómo se trabaja en el hito 2 y como se desarrollará la historia de usuario 5 y la historia de usuario 6.

**Planificación de hito3, Sprint 3:**

En la planificación que se mostrara a continuación, será una continuación de y un término a la historia de usuario 2, esto porque se fijaran los estándares de calidad apropiados para el proyecto, y en mutuo acuerdo con el product owner del proyecto, los entandares de calidad se fijaran una vez terminados las partes prácticas y funcionales del proyecto.



**Fase de inicio Hito 1**

A continuación de se programarán las tareas que ingresarán al hito para así dar confirmar las historias de usuario que se desarrollara y dar por iniciado el Sprint 1.

**La historia de usuario que ingresará al hito será la siguiente:**

**H.U 2**

* Como usuario de una plataforma big data quiero extraer y limpiar grandes volúmenes de datos para aplicar estándares de calidad a los datos a procesar.

La historia de usuario número 2 en el pasado seminario de título no fue terminada en su completitud, por cual entra como una historia de usuario con prioridad en el sprint 1 del seminario de título 2.

Las tareas que se desarrollaran en estas historias de usuario son las siguientes:

1.-Generar plan de proyecto para la historia de usuario

2.-Generar plan de gestión de riesgos para la historia de usuario medidas de “Contingencia” y “Mitigación”

3.-Planificación del hito y del sprint (Gestión de la configuración y gestión de cambios) con la herramienta GitHub trello.

4.-Generar las tareas que ingresaran como product backlog al sprint 1

5.-Generar métricas del comportamiento del trabajo de la historia de usuario

**Plan de entregables**

Cada Sprint terminará con un entregable que será liberado. Al final del proyecto el entregable será liberado al ambiente de producción de la universidad siendo instalado en servidores locales de la universidad para su posterior uso de procesamiento de datos.

Los entregables se comenzarán a emitir a partir del primer Sprint.

A continuación, se podrá observar los entregables para el Sprint 1, en donde se define que versión y documentos serán liberados al final de cada Sprint, y en qué ambiente estarán disponibles.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SPRINT | FECHA | VERSIÓN | DOCUMENTOS | AMBIENTE | RESPONSABLES |
| 1 | 06/0/2019 | V 1.0 | LANZAMIENTO DE MANUAL DE INSTALACION DE APACHE HUE | DESAROOLLO Y PRUEBAS | DAVID CÓRDOVA |
| 2 | 06/06/2019 | V 1.1 | LANZAMIENTO DE PLAN DE PRUEBAS DE SPRINT 1 | DESARROLLO Y PRUEBAS | DAVID CÓRDOVA |

**Gestión de la configuración.**

Para la gestión de la configuración del proyecto, esta se dividirá en dos partes, gestión de línea base y control de versiones, los cuales serán manejadas con la herramienta GitHub.

**Gestión de línea Base**

Cada vez que se dé por finalizado una etapa del proyecto será liberada una línea base del proyecto.

Cada Sprint que se dé por finalizado y este sea aprobado por el Scrum Master, se libera una línea base.

Para las presentaciones de proyecto y entrega de hitos, se anexará toda la información relacionada al hito y se generará una línea base del proyecto.

Cada línea base que posea el proyecto, será documentada en el repositorio local del ordenador y luego subido mediante COMMIT y PUCH a la plataforma GitHub.

A continuación, se podrá observar una imagen con el TAG en GitHub.

**Gestión de cambios para hito 1, historia de usuario 2**

El proceso de gestión de cambios será realizado tanto por la herramienta Trello, que nos permitirá gestionar tickets guardarlos y cerrarlos, y la plataforma GitHub, esta brindara más funcionalidades a la gestión de cambios, permite generar ticket cada vez que surjan problemas, cada ticket pertenecerá a un “Milestone” en el cual se define un periodo de tiempo, este se utilizara para el sprint que se ejecutara en este hito.

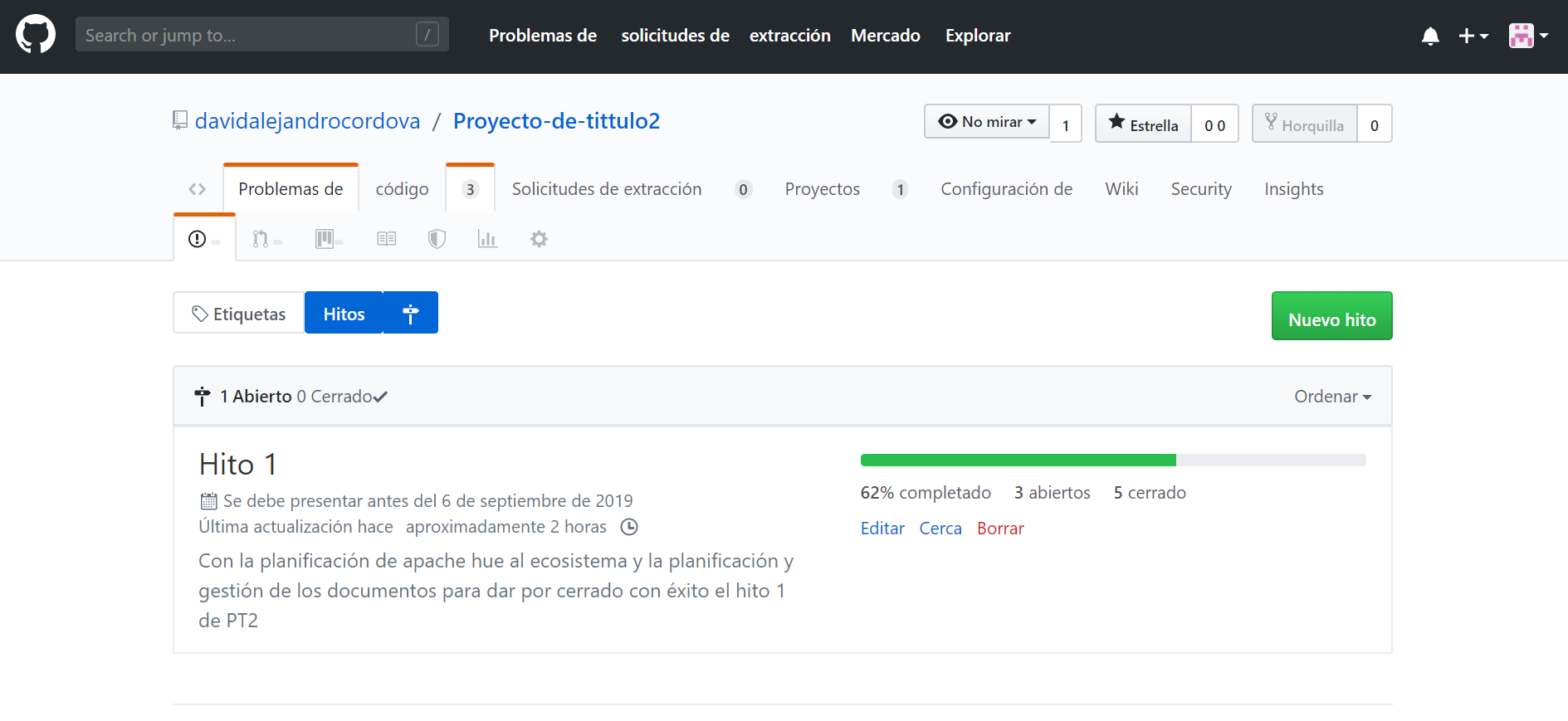
El Mileston 1 corresponderá a la fase de inicio del proyecto, el Milestone 2 al Sprint 1 del Hito 2 y el Milestone 3 al Sprint 2 del Hito 3.

El ciclo de vida para la creación de tickets son los siguientes:

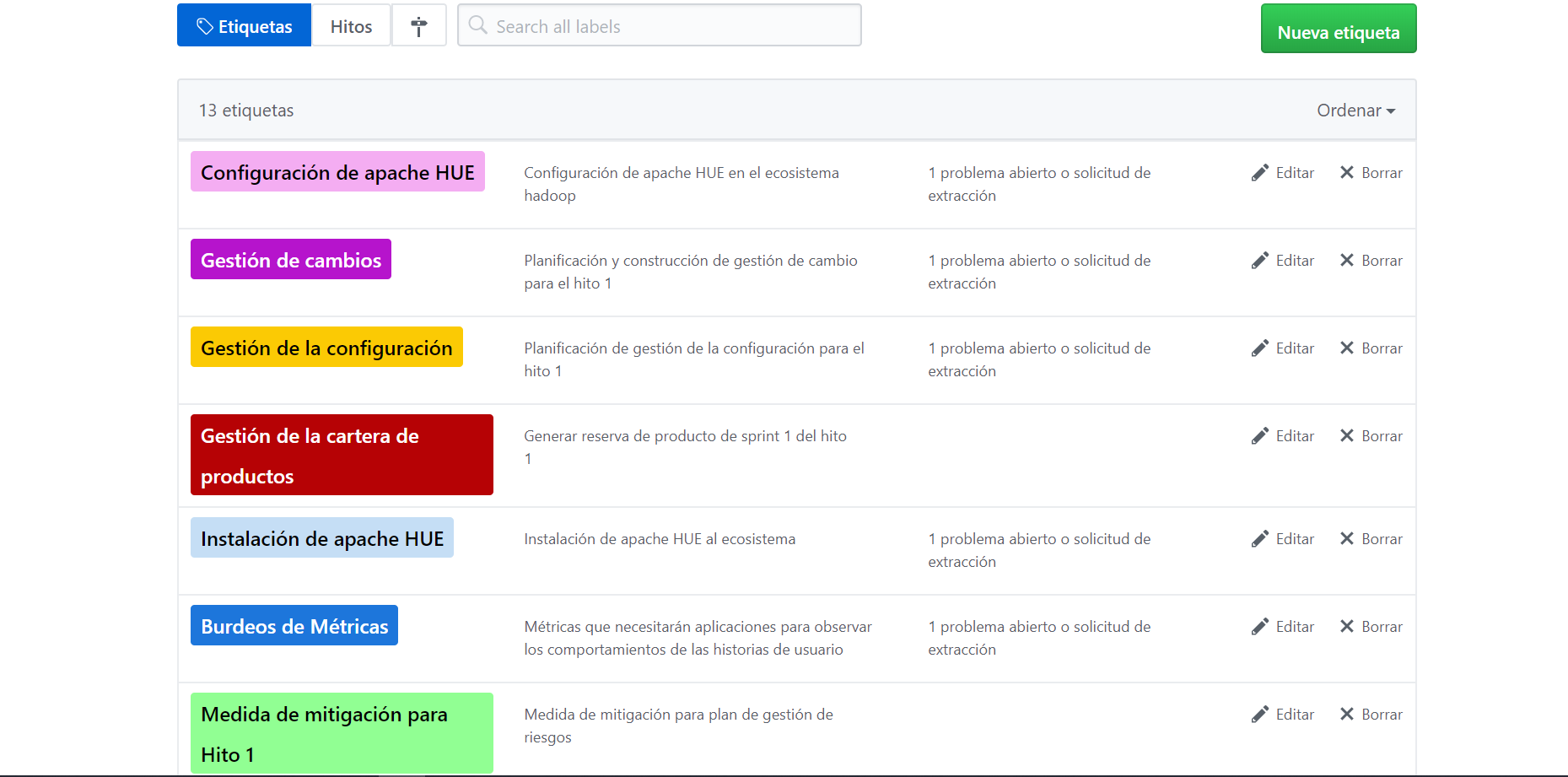
* Nuevos tickets: El ticket ha sido creado y especificado sus características
* Asignado a un desarrollador: El ticket ha sido asignado a un desarrollador (En este caso el equipo solo cuenta con 1 solo desarrollador)
* Cerrado: El ticket se cierra una vez resuelto
* Re-abierto: El ticket ha sido re-abierto por problemas sin solucionar

En la siguiente imagen se puede observar que los milestone de hito 1 se encuentra a un 62%, ya que aún existen tareas por terminar que no se han cerrado.

Se puede observar los hitos creados para el proyecto de título:

****

En la imagen que se podrá visualizar a continuación se mostrara como fueron realizados los lebel, los cuales cada uno de esto están asignado a los milestone.



A continuación, en la imagen que se mostrará se podrá observar como los milestone se encuentran relacionados con los lebel creados en la imagen anterior para conformar el hito 1 del proyecto de título.



Se puede visualizar en el rectángulo marcado en la imagen anterior que existen 3 milestone abierto que quedan por terminar y 5 cerrados los cuales fueron terminados con éxito.

Para acceder a la plataforma en donde se controla la gestión de cambio será mediante el siguiente link:

* **https://github.com/davidalejandrocordova/Proyecto-de-tittulo2.git**

**Control de versiones**

Todos los documentos creados serán controlados por versiones de desarrollo, los documentos que sean creados por primera vez serán asignados con la versión 1.0, como se muestra en las imágenes anteriores, cada vez que un documento sea modificado, a este nuevo documento con los cambios se genera una nueva versión, ejemplo “nombre v1.1” y así sucesivamente. En caso de crear nuevos documentos pasara a la versión 2.0.

Todas las versiones generadas se almacenarán en un repositorio local al interior de un ordenador, al cual se conectará con GitHub, el cual se encargará de llevar el registro de las versiones.

**Control de versiones para el código del proyecto**

El control de versiones para el código del proyecto, quiere decir que se realizaran versiones de los códigos internos para la instalación de apache HUE y las configuraciones pertinentes para su funcionamiento. Esto se realizará por motivos de pérdidas de documentos u archivos necesarios para la instalación o por problemas que puedan surgir en un futuro, esto permitirá volver a un estado anterior versionado y/o recuperar un estado deseado del ecosistema.

Para el control de versiones, se ha utilizado las herramientas de GitHub, tanto de plataforma para Windows que nos servirá como repositorio local, como la plataforma en la nube que nos servirá para visualizar los archivos de código en la nube para que terceros puedan ver estos archivos y gestionarlos si lo estiman conveniente.

**Fase de inicio de Sprint 1**

**Sprint 1**

**Tareas de la historia de usuario 2**

* Pruebas de ejecución de sistemas en Apache Hbase
* Investigación de compatibilidad de apache Hbase con apache HUE
* Proceso de instalación de apache HUE
* Proceso de acoplado e incorporación de apache HUE al ecosistema del proyecto

**Planificación de gestión de riesgos**

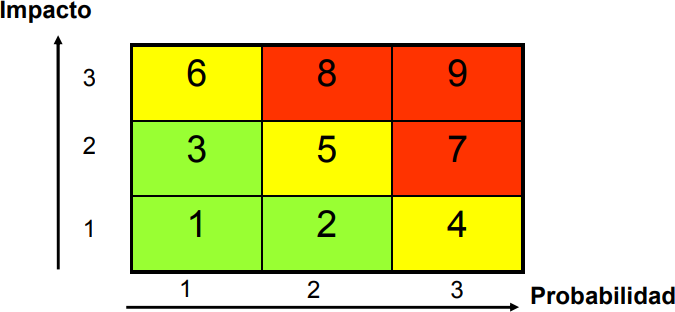
El proceso de gestión de riesgos consiste en la evaluación de todos los riesgos que afecten el desarrollo del proyecto semanalmente.

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. E este plan se concentra en las amenazas y se han clasificado en tres categorías:

**Riesgos Administrativos(RP):** Son aquellas situaciones cuyo disparador depende de circunstancias propiamente ligadas al área del proyecto.

**Riesgos Técnicos(RT):** Son aquellas situaciones cuyo disparador depende de circunstancias propiamente ligadas al área administrativa.

**Riesgos externos(RE):** Son aquellas actividades donde el disparador del riesgo no depende de los departamentos o áreas de la organización, sino de un externo como un proveedor de bienes o servicios.

Para lograr priorizar los riesgos identificados e utiliza una tabla de prioridad por impacto

**Porcentaje de amenaza = (Probabilidad \* impacto del riesgo) \*100**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ID | Riesgos | Probabilidad | Impacto | % de amenazas | Mitigación | Contingencia |
| RT1 | Desconocimientos de las herramientas a implementar | 0.5 | 0.5 | 25 | Estudiar las nuevas herramientas a implementar |  |
| RT2 | Problemas de compatibilidad sistemas operativo de Linux con las herramientas | 0.2 | 0.1 | 20 | Investigar problemas de compatibilidad de S.O con la herramientas a implementar | Cambiar herramienta que sea compatible con el S.O implementado |
| RT3 | Falta de conocimiento en codificación en Python de patrones y configuraciones | 0.6 | 0.7 | 42 | Estudiar la codificación en Python para la implementaciones patrones y configuraciones | Cambiar de lenguajes de programación para la codificación de patrones y las configuraciones |
| RT4 | Problemas en la instalación de la herramienta apache HUE | 0.5 | 0.6 | 30 | Estudiar la solución a los problemas de instalación de apache HUE | Cambiar de herramienta de apache a otras que tenga compatibilidad con Apache Hbase |
| RP1 | Fallo en el área técnica del proyecto | 0.8 | 0.8 | 64 | Dedicar más tiempo a encontrar las fallas del área técnica | Documentar y agregar área técnica al siguiente sprint. |
| RP2 | Problemas con las reuniones con el cliente | 0.5 | 0.3 | 15 | Estudiar la solución u/o cambiar día o lugar de las reuniones | Comunicarse mediante llamas telefónicas y/o mediante zomm (Llamadas en línea). |
| RE1 | Problemas con servidores de la universidad | 0.7 | 0.8 | 56 | Buscar forma de restablecimiento de los servidores de la universidad | Proyecto funcional en la nube. |
| RE2 | Problemas de robo del material de trabajo | 0.3 | 0.8 | 24 |  | Hacer uso de los respaldos en la nube de GitHub |

Riesgos que se presentaron en la realización de la historia de usuario 2

Se presentaron problemas en la instalación de la nueva herramienta apache HUE, la descarga de algunos paquetes de apache HUE no eran compatibles con la versión que se está ocupando en el proyecto, por lo cual se optó a la medida de mitigación, la cual consistía en la búsqueda y estudio de a solución.

El problema fue solucionado con éxito mediante el uso de la medida de mitigación planteada en la gestión de riesgos del proyecto

**Que es Data Mining**

La minería de datos, o data mining , es un campo perteneciente a las ciencias de la computación cuyos procesos intentan descubrir patrones en grandes volúmenes de conjuntos de datos. Entre otros, se utilizan métodos pertenecientes a los campos de inteligencia artificial, aprendizaje automático y estadísticas. El objetivo principal de Data Mining consiste en la extracción de información sobre un conjunto de datos y transformarla en una estructura intangible.

El proceso de data mining consta de los pasos detallados a continuación:

1. Selección del conjunto de datos
2. Análisis de las propiedades de los datos
3. Transformación de los conjuntos de entrada
4. Seleccionar y aplicar las técnicas de data mining
5. Extracción del conocimiento
6. Interpretación y visualización

Las instalaciones pendientes son apaches hue, el cual brindara facilidad con la entrega de una interfaz web que da soporte a apache hadoop y su ecosistema, Este nos entregara las siguientes herramientas que facilitaran el uso del ecosistema:

* Explorado de archivos de HDFS
* Job Browser para gestionar y visualizar tareas MapReduce y Yarn
* Explorador de la base de datos Hbase
* Editores de consultas para apache Hive, apache pig e impala
* Editor y gestión de apache Oozie
* Explorador de apache ZooKeeper
* Job Designer para crear tareas MapReduce
* Oozie Query editor para consultas PostGres, Mysql, Sqlite y Oracle

Data Mining

Para la minería de datos es necesario tener implementado todas las herramientas de apacha hadoop con las que se trabajaran y la plataforma weka, la cual es una plataforma de software para aprendizaje automático y data mining.

Los algoritmos que se implementarán serán:

* Algoritmo Kmeans: Consiste en concentrar grupos de datos similares o clusters
* Algoritmo FarthestFirst: Se considera como una variación de k-means ya que usa a misma función objetivo con la restricción de que los centros de los clusters deben pertenecer al dataset, su objetivo es determinar el mejor representante del centro de casa clauster.
* De entre las reglas de asociación se han seleccionado **A priori,** que se ejecuta de forma no distribuida, y **FPGrowth,** que realiza cálculos similares con el añadido de la paralización de sus operaciones.