Aldo Orozco Gomez Serrano

Enunciado del Alcance del Proyecto

[Adaptive Big Data pipelines]



# **Historial de Revisiones**

|  |  |  |  |
| --- | --- | --- | --- |
| Revisión # | Fecha de Revisión | **Descripción del Cambio** | **Autor** |
| 1 | 5/11/18 | Cambios iniciales | Aldo Orozco |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Tabla de Contenido

[**Historial de Revisiones** 1](#_Toc408847546)

[**1.** **Resumen Ejecutivo** 3](#_Toc408847547)

[1.1 Justificación 3](#_Toc408847548)

[1.2 Valor Agregado 3](#_Toc408847549)

[**2.** **Objetivos de Negocio** 3](#_Toc408847550)

[2.1 Objetivo 3](#_Toc408847551)

[2.2 Beneficios a Obtener 3](#_Toc408847552)

[2.3 Descripción del Producto a Obtener (solución) 3](#_Toc408847553)

[**3.** **Metodología del Proyecto** 3](#_Toc408847554)

[**4.** **Delimitación del Proyecto** 3](#_Toc408847555)

[4.1 Descripción del Alcance en Cada Etapa 3](#_Toc408847556)

[**5.** **Productos a Entregar** 3](#_Toc408847557)

[5.1 Productos a Entregar en Cada Etapa 3](#_Toc408847558)

[**6.** **Exclusiones** 3](#_Toc408847559)

[**7.** **Supuestos** 3](#_Toc408847560)

[**8.** **Restricciones** 3](#_Toc408847561)

[**9.** **Criterios de Éxito** 4](#_Toc408847562)

[**10.** **Factores de Éxito** 4](#_Toc408847563)

[**11.** **Riesgos** 4](#_Toc408847564)

[11.1 Principales Riesgos/Amenazas 4](#_Toc408847565)

[**12.** **Estimaciones Preliminares** 4](#_Toc408847566)

[12.1 Tiempo 4](#_Toc408847567)

[12.2 Costo 4](#_Toc408847568)

[**13.** **Matriz de Flexibilidad** 4](#_Toc408847569)

[**14.** **Información General** 4](#_Toc408847570)

[14.1 Acrónimos 4](#_Toc408847571)

# **Resumen Ejecutivo**

## Justificación

Actualmente la información es considerada como el recurso más importante para la mayoría de las empresas en crecimiento, debido a que facilita la toma de decisiones con base en hechos y permite conocer el comportamiento de los usuarios en tiempo real. En consecuencia, se recolecta cada vez más información para tratar de reaccionar rápidamente al mercado, lo que eventualmente resulta en sistemas ineficientes e inmantenibles para empresas pequeñas y medianas. Para solventar esto las empresas contratan ingenieros con conocimiento en Big Data para:

* Migrar la información de diversas fuentes a un entorno de nube
* Crear tuberías de extracción, transformación y carga de datos (ETL) a gran escala
* Crear APIs de consulta (salida) e ingreso de nuevos datos al sistema (entrada)

Desafortunadamente, debido a que el numero de tecnologías requeridas para la construcción suele ser alto (entre 5 y 20), se requiere un amplio *expertise* en las distintas tecnologías, lo que conlleva grandes perdidas de tiempo de desarrollo, y por consiguiente de dinero, para las empresas que desean obtener resultados de la información lo más pronto posible.

Para facilitar la migracion, transformacion y analisis de la informacion de forma eficiente y de forma semi-automatizada, se propone un sistema capaz de analizar y formar la infraestructura de nube deseada.

## Valor Agregado

Facilitar la migracion y generacion de tuberias de transformacion de datos, utilizando un esquema sencillo para personas o empresas no familiarizadas con las tecnologias de nube, otorgando a su vez codigo autogenerado eficiente, basado en la experiencia del uso de dichas tecnologias.

# **Objetivos de Negocio**

## Objetivo

Generar un producto de software capaz de generar valor a empresas que buscan obtener ventaja de la informacion que recolectan y que desean reaccionar de forma rapida al mercado, utilizando solo una fraccion del tiempo requerido en comparacion con la contratacion de personal altamente especializado, planeacion de infraestructura, desarrollo y optimizacion de la misma.

## Beneficios a Obtener

* Disminucion de costos de empresas para escalar sus analiticos
* Agilidad en la migracion de datos a la nube
* Visibilidad de metricos para reaccionar rapidamente a cambios en el mercado
* Facilidad de generacion de reportes

## Descripción del Producto a Obtener (solución)

Se propone un sistema que facilite:

* La migración de los datos locales de un grupo de fuentes conocidas (i.e., Excel, MySQL, Logs) a un entorno de nube (e.g., AWS o GCP)
* La creación de codigo adaptado a las transformaciones de la información a utilizar
* La creación de APIs de entrada y salida, para permitir el almacenamiento y extracción de datos de forma eficiente
* La generación de reportes básicos con base en filtros

El sistema proveerá una interfaz gráfica (GUI) para que el usuario proporcione los esquemas de entrada y salida de la tubería, así como las transformaciones básicas deseadas; en base a ello se determinarán los componentes y las configuraciones adecuadas.

Para efectos prácticos se considerarán 3 tipos de casos de uso para las tuberías: ventas, inventarios y eventos. Cada uno con su respectivo tipo de reporte.

# **Metodología del Proyecto**

Para este proyecto se propone una metodologia Agile, en particular la metodologia SCRUM que permite iterar rapidamente y generar pequeños entregables cada 3 semanas.

# **Delimitación del Proyecto**

## Descripción del Alcance en Cada Etapa

* Investigacion y prototipado
  + En la investigacion se revisaran un grupo de tecnologias actuales relacionadas al computo distribuido y la agrupacion/agregación de datos. De igual forma se investigará cual de las actuales plataformas de nube es la mas adecuada tanto para experimentación, como para empresas en la actualidad.
* Migracion a la nube
  + En esta etapa se implementara un artefacto de software que sea capaz de analizar la fuente de datos que el usuario seleccione y migre de forma eficiente a la plataforma de nube. Se contemplan 3 tipos de fuentes de datos de entrada: Logs (eventos), archivos CSV y bases de datos MySQL. Se limitara que ninguno de los archivos sea mayor a 30GB en tamaño.
* Creacion de tuberias de transformacion
  + En esta etapa se incluira la logica para que los archivos que se han subido a la nube se puedan manipular mediante una tuberia de extraccion, transformacion y carga (ETL). Esta logica generara optimizaciones en base a la secuencia de transformaciones especificadas. Dichas transformaciones resultaran en un almacenamiento en una capa de cache, tal como Apache Cassandra.
* Generacion de APIs
  + En esta etapa se incluira la logica para generar APIs que se podran consultar programaticamente para escribir en el data warehouse directamente, y para leer de la capa de cache generada anteriormente.
* Creacion de reportes
  + Dichos reportes de haran en Tableau y consistiran en visualizaciones basicas de la informacion agregada para poder realizar trabajo exploratorio. Se limitara a un par de graficas, tales como barras, pastel y de lineas.

# **Productos a Entregar**

## Productos a Entregar en Cada Etapa

* **Investigacion y prototipado:** Nada
* **Migracion a la nube:** Pieza de software inicial
* **Creacion de tuberias de transformacion:** Incremento de sistema para la generacion de las tuberias
* **Generacion de APIs:** Incremento de sistema para generar las APIs
* **Creacion de reportes:** Producto final de software capaz de generar visualizaciones basicas

# **Exclusiones**

* Guias de migracion para otras fuentes de datos
* Complejas transformaciones de datos como non-equi joins
* Fuentes de datos no estructurados
* Medidas de seguridad mas alla de las provistas por el proveedor de nube por defecto

# **Supuestos**

* La informacion estara en formato legible (texto, numeros) para hacer analisis de forma sencilla
* La computadora que contiene los datos a migrar debera ser capaz de conectarse a internet y tener un sistema operativo Linux o basado en Unix
* El usuario final tendra conocimientos intermedios acerca de como funcionan los sistemas distribuidos y las tuberias de transformacion

# **Restricciones**

* No existen garantias de tiempo de ejecucion de la tuberia debido a las variaciones que se pueden presentar en la data
* Los reportes son solamente visualizaciones de la informacion almacenada
* Medio a bajo volumen de datos, debido a restricciones de presupuesto
* Interfaz grafica sera sencilla y solo cubrira aspectos minimos funcionales

# **Criterios de Éxito**

* Las piezas de software son correctamente documentadas en codigo y probadas para casos de prueba basicos, los cuales tambien deben quedar documentados
* Asegurarse que la informacion quede intacta para consulta del usuario final
* Comunicación bilateral constante con Asesor y demas involucrados para garantizar la calidad del producto final
* Manejo de un tablero para tener control sobre las actividades realizadas

# **Factores de Éxito**

* Adecuada retroalimentacion por parte de asesor y demas involucrados
* Documentacion sencilla de encontrar y analizar para la creacion del software

# **Riesgos**

## Principales Riesgos/Amenazas

* Retrasos debido a falta de documentacion o errores aun no resueltos por parte del proveedor de servicios de nube
* Cambios de ultimo minuto

# **Estimaciones Preliminares**

## Tiempo

Se contemplan 4 semestres en total, de los cuales: el primero constituira la generacion de la idea y de la documentacion basica; el segundo cubrira la parte de investigacion y comparacion, asi como inicios de la implementacion; el tercero continuara con la implementacion y generara la documentacion correspondiente; el cuarto se avocara a refinar detalles de implementación finales y documentación para entrega del proyecto.

## Costo

El costo inicial que se prevee es del costo de la maestria (250,000 MXN), asi como 10,000 pesos para el prototipado en la nube (en caso de requerir mas poder de computo del otorgado por las pruebas gratuitas).

# **Matriz de Flexibilidad**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Alto | Medio | Bajo | Justificación |
| Tiempo |  |  | X | Debido a la complejidad del problema, es poco negociable el tiempo |
| Costo |  |  | X | Se prevee que no sera necesaria una inversion mucho mayor a la ya especificada debido al uso de herramientas open source |
| Alcance |  | X |  | Es posible ver una disminucion o cambio de alcance durante la ejecucion del proyecto, para garantizar que se obtenga la calidad y en tiempo |

# **Información General**

## Acrónimos

|  |  |
| --- | --- |
| **Acrónimo** | **Descripción** |
| ETL | Extraction, Transform and Load – etapa de transformacion y almacenamiento de información |
|  |  |
|  |  |
|  |  |
|  |  |