Google BigQuery

Luis Villalba 59191

Universidad Católica de Asunción, Departamento de Ciencias y Tecnologías,
Sede Santa Librada, Asunción, Paraguay
villalba.luifer@gmail.com

Abstract. En este paper estaremos adentrándonos en la herramienta que ofrece Google para manejo de datos a grande escala, datos que no pueden ser procesados convencionalmente, Google BigQuery.

Introducción

Qué es Google BigQuery? Google BigQuery es un servicio web de cloud computing. Es una herramienta bastante atractiva por su facilidad de uso, su funcionalidad y su precio. Ideal para aquellas empresas que no poseen la infraestructura necesaria para procesar una gran cantidad de información.
El servicio web de Google BigQuery permite realizar el almacenamiento y consulta de conjuntos de datos masivos con billones de filas. Su uso es sencillo y permite a los desarrolladores y analistas de negocios estudiar bases de datos en un tiempo casi real. Orientado directamente al análisis de datos, es decir si vamos al marco conceptual estamos hablando de una solución con orientación OLAP (On-Line Analytical Processing).

En siguiente documento tiene como objetivo el de conocer esta herramienta con unos previos conceptos antes de tocar el tema.

1 Big Data

Qué es Big Data y porqué se ha vuelto tan importante? Pues bien, en términos generales podríamos referirnos como a la tendencia en el avance de la tecnología que ha abierto las puertas hacia un nuevo enfoque de entendimiento y toma de decisiones, la cual es utilizada para describir enormes cantidades de datos (estructurados, no estructurados y semi estructurados) que tomaría demasiado tiempo y sería muy costoso cargarlos a un base de datos relacional para su análisis. De tal manera que, el concepto de Big Data aplica para toda aquella información que no puede ser procesada o analizada utilizando procesos o herramientas tradicionales. Sin embargo, Big Data no se refiere a alguna cantidad en específico, ya que es usualmente utilizado cuando se habla en términos de petabytes y exabytes de datos. Entonces Cuánto es demasiada información de manera que sea elegible para ser procesada y analizada utilizando Big Data? Analicemos primeramente en términos de bytes:
Gigabyte = $10^9 = 1,000,000,000$
Terabyte = $10^{12} = 1,000,000,000,000$
Petabyte = $10^{15} = 1,000,000,000,000,000$
Exabyte = $10^{18} = 1,000,000,000,000,000,000$
Además del gran volumen de información, esta existe en una gran variedad de datos que pueden ser representados de diversas maneras en todo el mundo, por ejemplo de dispositivos móviles, audio, video, sistemas GPS, incontables sensores digitales en equipos industriales, automóviles, medidores eléctricos, veletas, anemómetros, etc., los cuales pueden medir y comunicar el posicionamiento, movimiento, vibración, temperatura, humedad y hasta los cambios químicos que sufre el aire, de tal forma que las aplicaciones que analizan estos datos requieren que la velocidad de respuesta sea lo demasiado rápida para lograr obtener la información correcta en el momento preciso. Estas son las características principales de una oportunidad para Big Data. Es importante entender que las bases de datos convencionales son una parte importante y relevante para una solución analítica. De hecho, se vuelve mucho más vital cuando se usa en conjunto con la plataforma de Big Data. Pensemos en nuestras manos izquierda y derecha, cada una ofrece fortalezas individuales para cada tarea en específico. Por ejemplo, un beisbolista sabe que una de sus manos es mejor para lanzar la pelota y la otra para atraparla; puede ser que cada mano intente hacer la actividad de la otra, mas sin embargo, el resultado no será el más óptimo. **[1]**

2 Cloud Computing

Cloud computing es el desarrollo y la utilización de capacidad de procesamiento computacional basado en Internet (la nube). El concepto es un cambio de paradigma, a través del cual los usuarios ya no necesitan contar con conocimientos, experiencia o control sobre la infraestructura tecnológica que se encuentra en la nube, la misma que soporta sus actividades. Este concepto involucra típicamente la provisión de recursos fácilmente escalables y casi siempre virtualizados, tratados

como servicios sobre Internet.
El termino nube (*cloud* en ingles) es usado como una metáfora para el Internet, basado en como el Internet es representado en los diagramas de redes computacionales y como abstracción de la infraestructura subyacente que el misma oculta. Los proveedores de cloud computing proveen aplicaciones en línea de negocio, las mismas que se pueden acceder desde exploradores de internet (Firefox, IE, Opera, Chrome, Safari, etc.), mientras el software y los datos son almacenados en los servidores. [**2**]

3 Inicios de BigQuery

Big Data está convirtiendo en una parte integral de la logística de negocio de la empresa. Hadoop es el marco y la tecnología detrás de Big Data.
Ofrece diversas herramientas para ingerir, procesar y analizar grandes conjuntos de datos que normalmente se ejecutan en unos pocos terabytes de tamaño.
Aunque Hadoop ha madurado, sigue siendo considerado para el procesamiento por lotes. La consulta y el análisis de datos en tiempo real con Hadoop es difícil y costoso.
El corazón de Hadoop es el MapReduce, que no es el adecuado para consultas interactivas. Las tecnologías como Impala y Apache Spark de Cloudera comenzaron a complementar MapReduce para hacer frente a los datos en tiempo real.
Aunque era Google el que en gran medida contribuyó al paradigma de MapReduce, también es uno de los primeros en identificar sus inconvenientes. Los ingenieros de Google se dieron cuenta de que MapReduce no era ideal para consultar conjuntos grandes, distribuidas de datos en tiempo real. Para solucionar este problema, Google desarrolló una herramienta interna denominada Dremel, que permitió ejecutar consultas SQL en grandes conjuntos de datos en tiempo real.
Dremel ha sido diseñado para ofrecer un excepcional rendimiento de consulta rápida sobre los conjuntos de datos de datos distribuidos que se almacenan a través de miles de servidores. Es compatible con un subconjunto de SQL para consultar y recuperar datos.
En Google I/O 2012, Google anunció BigQuery que expuso Dremel con el mundo exterior como un servicio en la nube. Desde entonces, BigQuery ha evolucionado hasta convertirse en un alto rendimiento y escalable motor de consulta en la nube.

4 Características

Algunas carácteristicas que posee esta herramienta:

- **Velocidad**, puede analizar miles de millones de filas en segundos.
- **Escalable**, capacidad de manejo de un gran tamaño de datos. Miles de millones de registros con un tamaño de terabytes de datos.
- **Simple**, lenguaje de consulta similar a SQL. Alojado en la infraestructura de Google.

- **Múltiples permisos**, diferentes accesos de usuario dependiendo de los permisos que se otorguen.
- **Seguridad**, posee acceso **SSL** (Secure Sockets Layer).
- **Múltiples métodos de acceso**, conectarse a BigQuery utilizando el navegador BigQuery, la herramienta de línea de comandos bq, API o Google Apps Script.

5 Fundamentos de BigQuery

5.1 Proyectos

Los proyectos son contenedores de alto nivel en la plataforma de la nube de Google. Almacenan información sobre la facturación y los usuarios autorizados, y contienen datos BigQuery. Cada proyecto tiene un nombre y un identificador único.

5.2 Tablas

Las tablas son las que contienen los datos en BigQuery, junto con el esquema de la tabla correspondiente que describe los nombres de campos, tipos y otra información.
BigQuery soporta vistas, tablas virtuales definidos por una consulta SQL. Además crea tablas en una de las siguientes maneras:

- Cargando datos en una nueva tabla.
- Ejecución de consultas.
- Copiando una tabla.

5.3 Datasets

Datasets, es un conjunto de datos así como su nombre lo indica, corresponde a los contenidos de una única tabla de base de datos en su versión más simple.
Los datasets permiten organizar y controlar el acceso a las tablas. Debido a que las tablas están contenidas en bases de datos, se tendrá que crear al menos un dataset para cargar datos en BigQuery.

5.4 Jobs

Los jobs, o trabajos, son acciones que se construyen y BigQuery ejecuta para cargar datos, exportar datos, consultar de datos, o copiar datos.
Ya que los trabajos pueden tardar mucho tiempo en completarse, se ejecutan de forma asincrónica y se puede consultar su estado.
BigQuery guarda la historia de todos los trabajos asociados al proyecto, accesible a través de la consola para desarrolladores de Google.

6 Interactuando con BigQuery

Hay tres principales maneras de interactuar con BigQuery.

6.1 Cargando y exportando datos

Antes de que se pueda hacer consultas, primero habría que cargar los datos en BigQuery. Si desea obtener los datos de nuevo de BigQuery, se debe exportar los datos.

6.2 Consulta y visualización de datos

Una vez que usted carga sus datos en BigQuery, hay algunas formas de consulta o vista de los datos de las tablas, estaremos citando algunas:

Consulta de datos

- Llamando al método *bigquery.tabledata.query()*.
- Llamando al método *bigquery.tabledata.insert()* con una configuración de consultas.

Visualización de datos

- Llamando al método *bigquery.tabledata.list()*.
- Llamando al método *bigquery.jobs.getQueryResults()*.

6.3 Manejo de datos

Además de la consulta y visualización de datos, puede administrar los datos en BigQuery utilizando funciones que permiten las siguientes tareas:

- Listar proyectos, jobs, tablas y datasets.
- Obtener información sobre jobs, tablas y datasets.
- Actualización de tablas y datasets.
- Borrar tablas y datasets.

7 Precio

BigQuery utiliza una estructura de datos de tipo columnar. Se le cobra al usuario el total de datos procesados en las columnas que seleccione, y el total de los datos por columna se calcula basándose en los de tipos de datos de la columna.
BigQuery ofrece opciones de precios escalables y flexibles para poder adaptarse a cualquier proyecto y presupuesto.
BigQuery cobra por almacenamiento de datos, pero para cargar y exportar datos no tiene costo adicional.

7.1 Operaciones sin costo

Las operaciones que se encuentran libre de todo costo son:

- Carga de datos.
- Exportación de datos.
- Lectura de tabla.
- Copia de tabla.

7.2 Precio de almacenamiento

El precio de almacenamiento esta basada en el tamao de datos almacenados en las tablas, basados en el tipo de datos que es almacenado.
El precio de almacenamiento es prorrateado por MB/segundo. Por ejemplo, si se almacena 500MB para la mitad de un mes, se debería pagar $0.0065.
El precio de almacenamiento es de $0.026 por GB, por mes. [**6**]

7.3 Precio de consultas

BigQuery ofrece dos opciones de precios: por demanda y por capacidad reservada.

Precio por demanda Las consultas por demanda utilizan un *pool-shared* de recursos entre los usuarios, a diferencia de los precios por capacidad reservada. El primer TB de dato procesado en el mes, las consultas que retornan errores y las consultas guardadas en cache no tienen costo.
Las consultas se cobran $5 por TB. [**6**]

Precio de capacidad reservada Para mayores cargas de trabajo, más consistentes, donde el precio va a de un throughput de 5 GB por segundo con un precio de $20.000 por mes. [**6**]

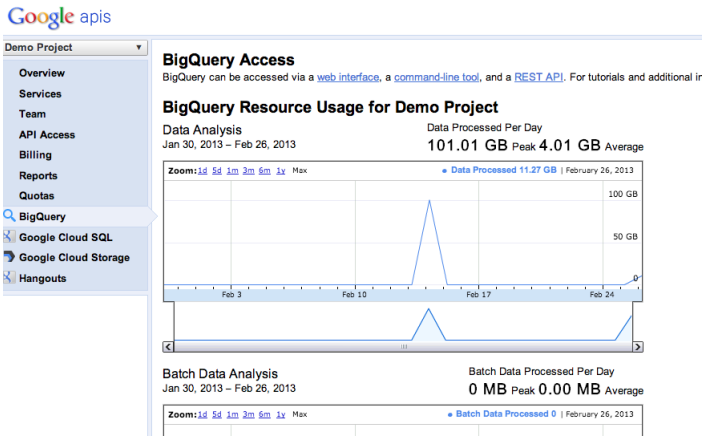
8 Uso

Una vez que haya iniciado sesión para BigQuery, se puede acceder al servicio de diferentes formas:

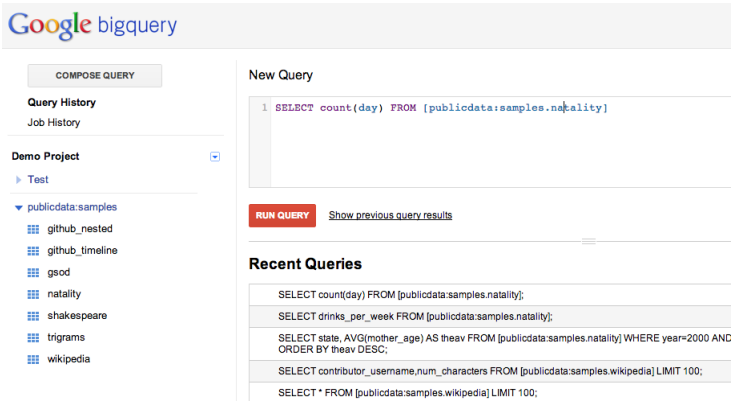
- **BigQuery Browser Tool:** con esta herramienta se puede facilmente navegar o crear tablas, correr consultas y exportar datos al servicio de almacenamiento en la nube de Google.
- **bq Command-line Tool:** usar la herramienta command-line de Python para el manejo y consultas de los datos.
- **REST API:** utilizando la API basada en REST para acceder a BigQuery mediante programación.

9 Ejemplo de Uso

Bigquery forma parte de las APIs que tiene disponible Google desde la consola de servicios.(Google API console).



Como muestra la figura, podemos activar el servicio y tendremos la posibilidad de administrar Bigquery por medio de una línea de comandos o una interface Web. Veamos la parte Web de administración.



Como pueden observar en la Consola de administración Web de Bigquery podemos generar nuevos proyectos o bien utilizar consultas de prueba en los repositorios públicos como natality, trigrams, etc. [3]

10 Uso en pequeñas empresas

BigQuery, es esencialmente un servicio online orientado a procesar grandes volúmenes de datos e información, es un servicio orientado a un cliente profesional, por tanto

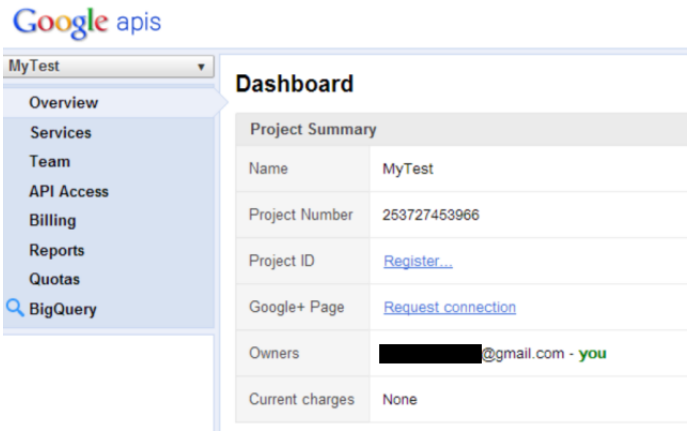
estamos hablando de un producto orientado a empresas.
Resumiendo, toda la experiencia y conocimiento que Google posee sobre almacenamiento masivo, gestión y proceso de datos, lo pone a disposición de empresas que no tienen la opción de disponer ni de la tecnología, ni de la arquitectura, ni de los recursos necesarios.
En un tiempo en el que las empresas necesitan ser tan competitivas y que el analizar y convertir el dato en información significativa de negocio resulta tan importante, en un tiempo en el que auge del Business Intelligence es una realidad y en el que ya se oye demasiado el concepto Big Data, esta puede ser una herramienta muy útil para que las empresas empiecen a experimentar y a adentrarse, a precios razonables, en las posibilidades que pueden ofrecer sus datos y los de los demás para mejorar los resultados actuales.

11 Y qué no es Google BigQuery?

Google BigQuery no es un sistema de reporting, no dispone de una interfaz en la que se pueda explotar de manera directa la información que se vaya obteniendo. Por lo tanto, lo ideal sería exportar los resultados obtenidos y volcarlos en otros sistemas de visualización con los que los usuarios que explotarán la información ya estén familiarizados (Tableau, Gephi, etc.).

12 Overview

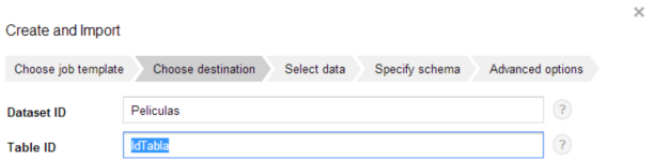
En este FrameWork ya podremos crear nuestros proyectos y activar los servicios que consideremos.



13 Crear Tabla

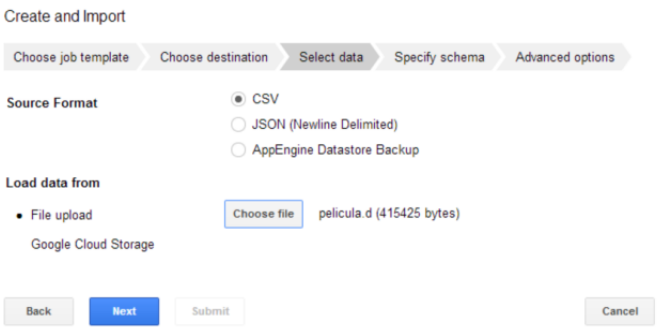
Una vez tenemos definido todo nuestro espacio de trabajo, ya podemos proceder a la importación de datos.

Para subir los datos a la nube habrá que crear un DataSet, una tabla, definir el schema de la tabla y hacer un upload de los ficheros necesarios.
Dentro del conjunto de datos (Dataset) informamos la tabla.



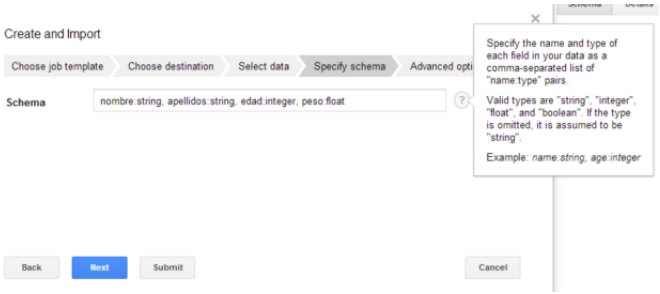
14 Selección de fichero

Deberemos seleccionar el fichero que contiene los datos que queremos subir a la nube.

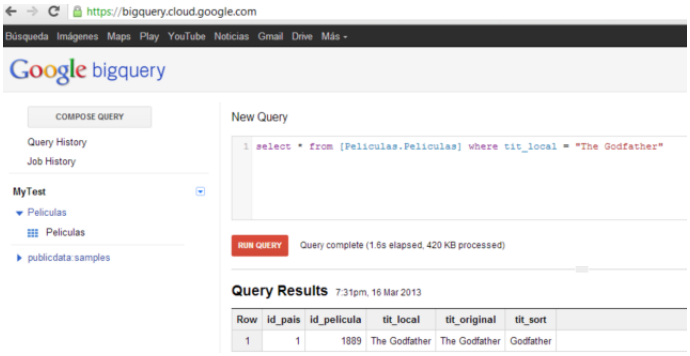


15 Definición del Schema

Finalmente definiremos el formato de la tabla informando los campos con el siguiente formato:
campo:formato, campo:formato, campo:formato



Los datos han sido subidos al Cloud y ya pueden ser explotados por BigQuery. Hay que recordar que la subida de datos y la posterior consulta viene implementada con la API propia de BigQuery y mediante Webservice REST. Todo un lujo para quienes quieran crear todo tipo de integraciones con otras aplicaciones.
Un ejemplo de una consulta desde la consola es el siguiente:



16 Impacto en las empresas

Si tenemos en cuenta que muchas empresas que se adentran en el mundo del Big Data suelen recurrir a consultorías que no son precisamente baratas, gracias a BigQuery las empresas pueden hacer sus primeros pinitos en el procesamiento de grandes volúmenes de información aunque, seguramente, si quieren hacer algo en gran profundidad y exprimir los datos al máximo acabarán recurriendo a los servicios de un tercero.
Conforme vaya aumentando la necesidad de las empresas por exprimir los datos que manejan y, por tanto, profundizar mucho más en el conocimiento de su sector, sus clientes o las oportunidades de negocio, supongo que irán surgiendo más iniciativas vinculadas a extender el uso del Big Data en más y más sectores productivos. Empresas como Google llevan haciendo esto desde hace bastantes aos aunque lo hacían de manera interna, por tanto, creo que aquí tienen un nicho de mercado bastante suculento en el que podrían hacerse un hueco y ofrecer un servicio llave en mano (infraestructura en la nube y software) con el que podría ser complejo competir.

17 Google BigQuery vs MapReduce vs PowerDrill

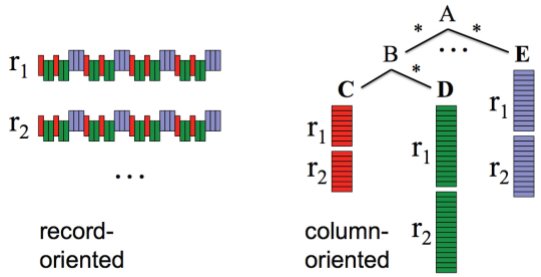
Antes de continuar con la comparación vamos a definir los tres:

- **Dremel**, es un servicio de consulta que le permite ejecutar consultas SQL en enormes conjuntos de datos. Es muy sencillo de usar y una gran experiencia de desarrollo no es requerida para su uso.
- **Apache Drill**, es similar a Dremel y BigQuery. Sin embargo, es una versión de código abierto y es bastante flexible, ya que soporta muchos mas lenguajes de consulta y formato de datos/fuentes. La diferencia que tiene con BigQuery es que PowerDrill mantiene los datos frecuentemente en memoria.

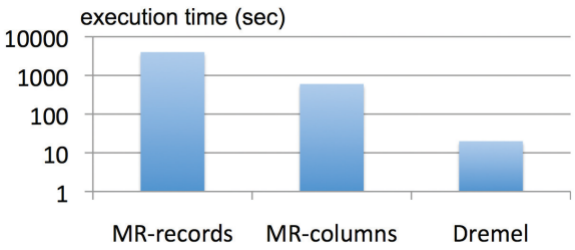
17.1 Qué hace a BigQuery tan rápido?

BigQuery puede escanear cientos de millones de filas no indexadas en menos de un minuto. Hay dos cosas que lo hacen tan rápido: Columnar Storage y Tree Architecture.

- **Columnar Storage:** en vez de mirar a los datos en forma de filas, los datos son almacenados como columnas. La ventaja de este tipo de almacenamiento es elevada con respecto al almacenamiento orientada a filas. Se analizan solo los valores necesarios, esto significa que solo 5-20 columnas tienen que ser visitadas de las miles disponibles, esto reduce considerablemente la latencia.
- **Tree Architecture:** se utiliza para el procesamiento de consultas y para añadir los resultados entre los diferentes nodos. BigQuery se extiende por miles de servidores mediante el uso de un árbol binario, los datos se encuentran fragmentados en varios equipos. Esto ayuda a recuperar los datos de una manera mas eficiente. [8]



17.2 Cuál es la diferencia entre MapReduce y BigQuery?



La principal diferencia entre MapReduce y BigQuery es que MapReduce se utiliza para procesar conjuntos de datos mientras que BigQuery se utiliza para analizarlos.
MapReduce procesa lotes de grandes volúmenes de datos y puede tardar horas o incluso días en hacerlo. Si se comete un error, el proceso debe reiniciarse. Se utiliza generalmente para *data mining* donde grandes conjuntos de datos no estructurados deben ser analizados mediante programación.
Las consultas de BigQuery generalmente terminan en menos de un minuto. Por

lo tanto, los métodos de ensayo y error son válidos a la hora de ejecutar consultas. Se utiliza generalmente para *Analytical Processing / Business Intelligence* donde grandes archivos CSV estructurados están disponibles. [**8**] En resúmen, mientras más sea la cantidad de pedazos son mas rápidas las consultas.

18 Cuando usar Google BigQuery?

La fuerza de BigQuery reside en su capacidad para manejar grandes conjuntos de datos. Por ejemplo, la consulta de decenas de miles de registros puede tardar sólo unos segundos. Incluso después de dos veces el número de registros, BigQuery tomaría el mismo tiempo para procesar la consulta. Dado que se basa en el lenguaje de consulta SQL estándar, es bastante fácil de construir consultas complejas para recuperar los datos. BigQuery es un almacén de datos estructurados en la nube. De ello se sigue el paradigma de tablas, campos y registros. Sin embargo, a diferencia de RDBMS, BigQuery admite campos que pueden contener más de un valor por lo que es fácil de consultar datos anidados repite. Google replica los datos BigQuery través de múltiples centros de datos para que sea de alta disponibilidad y duradero.
BigQuery normalmente se produce al final de la tubería de Big Data. No es un reemplazo a las tecnologías existentes, sino que los complementa muy bien. Después de procesar los datos con Apache Hadoop, el conjunto de datos resultante puede ser ingerido en BigQuery para su análisis. Corrientes en tiempo real que representan los datos del sensor, los registros del servidor web o gráficos de medios de comunicación social pueden ser ingeridos en BigQuery que se debe consultar en tiempo real. Después de ejecutar los trabajos de ETL en RDBMS tradicionales, el conjunto de datos resultante se puede almacenar en BigQuery. Los datos pueden ser ingeridos a partir de los conjuntos de datos almacenados en Google Cloud Storage, a través de la importación directa de archivos oa través de streaming de datos. Así que, si Apache Hadoop es el medio para Big Data, BigQuery es el final.

Conclusión

Lo primero que uno se pregunta cuando ve esta herramienta podría ser: Para que puedo utilizar esta tecnología? Todo el análisis de datos del tsunami tecnológico actual que se podría realizar, una mina de oro!.
Algunos ejemplos de utilización pueden ser: reportes entandarizados, análisis, exploración y minería de datos concreta desde diferentes clientes de acceso. Actualmente, la cantidad de información que se debe procesar es cada vez mayor, por ende el tiempo que lleva gestionarlos también va en aumento. Esto hace que la capacidad de software habitual no sea suficiente para el manejo de cantidades de datos muy grandes.
Aquí es donde aparece el nuevo término denominado Big Data. Un ejemplo sería una Base de Datos estructurada, que mientras crece la cantidad de datos guardados, la misma va perdiendo eficiencia. Por eso van a apareciendo lo que se denomina Bases de Datos no estructuradas, las cuales buscan mantener una eficiencia constante a medida que aumenta la cantidad de información guardada.
Por lo tanto, para manejar grandes cantidades de datos casi en tiempo real sin poder tener la infraestructura necesaria, aparece Google BigQuery. Es una propuesta de google con precio accesible.
Con el manejo de datos no estructurados hay una infinidad de usos que se le podría dar, información tan valiosa que está esperando ser explotada, como estudios de lugares frecuentados por personas a determinadas horas del día, donde frecuentas las barreras policiales (que calles frecuentan, la hora, días de la semana, etc.). Solo debería estar disponible una gran cantidad de información útil y la mejor forma de analizar dicha información. No hay que preocuparse de infraestructura ni de manejar nodos como es en el caso de Hadoop, ya que Google BigQuery ya se ocupa de eso.
Gracias a este trabajo pude ver la mina de oro que hoy día es la información disponible en la red, y las diferentes opciones para poder manejarla además de Google BigQuery. Pero una de las opciones más actractivas es la opción que nos propone Google para los que no tienen la posibilidad de tener un clúster de computadoras debido a su alto costo.

Referencias

1. Qué es Big Data?
   www.ibm.com/developerworks/ssa/local/im/que-es-big-data/
2. Cloud Computing
   http://cloud-america.com/?page_id=257
3. Moderno enfoque para trabajar datos masivos
   http://programa-con-google.blogspot.com/2013/03/moderno-enfoque-para-trabajar-datos.
   html
4. Google BigQuery. Consultas de alto rendimiento.
   http://ismasantacruz.wordpress.com/2013/03/19/google-bigquery-consultas-de-alto-rendi
5. Google Analytics Premium se integra con BigQuery
   http://trucosgoogleanalytics.com/google-analytics-premium-se-integra-con-bigquery-29
   #axzz3CIhRCrCE
6. Google BigQuery Documentation
   https://developers.google.com/bigquery
7. When to use Google BigQuery?
   http://cloudacademy.com/blog/when-to-use-google-bigquery/
8. BQ vs MR vs PowerDrill
   http://geeksmirage.com/google-bigquery-vs-mapreduce-vs-powerdrill/