# Instituto Tecnológico de Costa Rica

## Proyecto #2

*Alejandro Rojas*
*Saúl Zamora*

profesor
Kevin Moraga

June 20, 2017

# 1 Introducción

Se requiere un sistema para una empresa productora de dispositivos de IoT (Internet Of Things), los cuales se distribuyen a nivel nacional y se exporta a Nicaragua, Panamá y Guatemala.

Uno de los principales negocios de la empresa son las "Smart Houses", lo cual implica brindarle al cliente una experiencia única al darle control total de su casa y los dispositivos en ella.

Las bases de datos deben almacenar información de forma local y otros datos de forma remota en las instalaciones de la empresa. Actualmente, las tecnologías usadas son MySQL, SQL Server y Oracle.

La base de datos de producción posee un inventario de todas las partes que la empresa adquiere como materia prima para la creación de los dispositivos. Existe otro sistema que usa otra base de datos donde se controlan las ventas de los productos y dispositivos, la cual es básicamente un catálogo de los productos; además de un historial con los cambios en los precios de dichos productos.

Cada casa de habitación cuenta con una base de datos local, cuya función principal es recolectar información post-venta de los dispositivos e información generada por su uso.

Es requerido mantener las bases de datos locales y remotas replicadas para brindarle al usuario la experiencia que desea.

Además, el sistema deseado debe ser capaz de producir estadísticas sobre inteligencia de negocios con el fin de ayudar a la gerencia a tomar decisiones más acertadas. Algunos de los puntos más relevantes son:

- Costos de producción
- Sugerencias de consumo de productos del cliente en los próximos 15 días.
- Dispositivos IoT más utilizados por un cliente
- Dispositivos IoT más vendidos
- Tiempo de entrega de materiales relacionados a los dispositivos IoT más vendidos
- Estadísticas sobre los proveedores (tiempos de respuesta, calidad de materiales, etc)
- Determinar los materiales de mayor y menor circulación
- Estadísticas sobre las ventas anuales
- Estadísticas sobre los carriers de los productos (volumen enviado, más utilizados, destinos, etc)
- Índices de ganancias vs índices de gastos

# 2 Ambiente de desarrollo

Se utilizaron las siguientes herramientas para la elaboración del proyecto:

- Sistema operativo host utilizado: Windows 10
- Python 3.6 (Simulación)
- Oracle VirtualBox 5.1
- Windows Server 2008 R2 (Dos máquinas virtuales)
- SQL Server 2008 Enterprise
- Pentaho 7.1

# 3 Estructuras de datos usadas y funciones

## 3.1 Simulación del estado actual del sistema

En primera instancia tenemos el requisito de simular las entradas y salidas de datos de la compañía, correspondientes a los últimos 5 años; teniendo en mente que se dichos datos deben permitir realizar las consultas de Inteligencia de Negocios solicitadas por el gerente. La compañia cuenta con 3 fuentes de datos:

- BD de Producción: Posee un inventario de todas las partes que la empresa compra como materia prima para crear sus dispositivos; se pueden ver todas las ordenes de compra que se hacen a sus proveedores así como cuando dicha orden arriba a las bodegas de la fabrica. Toma en cuenta aspectos como devoluciones, defectos de partes, reparación entre otros. En las ordenes de compra es posible ver los precios que son facturados y los que financiero termina cancelando a los proveedores.
- BD Ventas: Controlan las ventas de los productos o dispositivos como se les quiera llamar, básicamente lo que se tiene es un catálogo de productos con su nombre, número de parte, descripción, los nombres de los manuales del producto y la ruta donde se encuentra los archivos digitales, así como informacón complementaria de los manuales. También toda la historia de precios del producto según una fecha en particular.
- Excel de Despachos: Se utiliza cuando un producto se despacha para ser distribuido ya sea a nivel nacional o internacional se hace a través de algún carrier o distribuidor utilizando una guá de envó. Existen chequeadores que anotan en una hoja de Excel: nombre del chequeador, fecha, hora, nombre del carrier, número de guía, cantidad del producto despachada, país destino y consumidor.

Teniendo en cuenta estas tres definiciones se desarrollaron los diseños de cada base de datos, y en base a estos se implementó un script en Python 3.6, el cual genera los datos aleatorios necesarios. Como estructura principal se utilizaron listas para almacenar los datos aleatorios, los cuales son modificados a medida que se simulan los 5 años de entradas y salidas, y durante este proceso se van guardando los distintos datos en archivos CSV. La simulación genera todos los archivos en una sola corrida, por lo que no utiliza otras funciones en particular. Sin embargo, son necesarias las siguientes librerías para su correcta ejecución:

- csv
- random (randint)
- datetime (date, timedelta)
- operator
- numpy

Cada CSV generado corresponde a una tabla en las bases de datos, en las figuras 1 y 2 ubicadas en los Anexos, se puede visualizar los diagramas para las bases de Inventarios y Ventas, respectivamente.

## 3.2 Replicación de Bases de Datos

Como segundo requisito es necesario implementar un esquema de replicación de forma que la base de datos de inventario se replique en la de productos y la de productos se replique en la de inventario. Para lograr este objetivo se utilzaron los *Replication Services* ofrecidos por SQL Server 2008 R2, esto consiste en crear *Publishers*, los cuales publican incialmente una instantánea de la base de datos (o *Snapshot*), y *Subscriptors*, los cuales utilizan esta instantanea para replicar el esquema y datos iniciales publicados. Una vez que los subcriptores tienen el esquema inicial, cada modificacion subsiguiente se maneja de manera transaccional, esto es, que los publicadores unicamente envían cuales fueron los cambios en particular que se realizaron.
En la Figura 4, en los Anexos, se puede visualizar ambas bases de datos y sus respectivas replicaciones, así como los publicadores y subscriptores mencionados.

## 3.3 Inteligencia de Negocios

Finalmente, una vez disponibles todas las fuentes de datos, se deben implementar las consultas relevantes tanto del estado actual del negocio, así como información historica que puedan ser de interes para los gerentes de la compañia. Para este objetivo se utilizó la herramienta Pentaho, la cual permite realizar tareas de Extracción, Transformación y Carga de datos (*ETL* por sus siglas en ingles), esto nos permite unir tablas y datos de las diferentes fuentes, y utilizar una sola fuente de información al realizar las consultas. En la Figura 3, en los Anexos, podemos ver una configuración para conectarse a una base de datos

SQL Server.

Este proceso involucra la creación de Cubos OLAP, *OnLine Analytical Processing*, los cuales son bases de datos multidimensionales, estos permiten analizar, por ejemplo, datos financieros por producto, por perodo, por país, por tipo de ingresos y de gastos, etc. Estos parmetros en función de los cuales se analizan los datos se conocen como *dimensiones*. Para acceder a los datos slo es necesario indexarlos a partir de los valores de las dimensiones o ejes.
Adicionalmente, Pentaho permite publicar dichos cubos a un servidor web, el cual nos permite generar reportes en base a estos cubos, manipulandolos a necesidad para mostrar los datos que el usuario necesite.

# 4 Instrucciones para ejecutar el programa

En lo respectivo a la simulación, para generar los archivos CSV, basta colocar el archivo *dataSimulator.py* en la carpeta donde se desee almacenarlos. Este se puede ejecutar ya sea mediante el IDLE de Python, presionando F5, o mediante la línea de comandos, ejecutando directamente el programa. Una vez generados los archivos, se puede utilizar el servicio de *import* de SQL Server para insertar los datos en las bases de datos respectivas.

Una vez cargados los datos, para visualizar las consultas y reportes, debemos ejecutar el archivo *start-pentaho.bat*, ubicado en la carpeta de instalación de Pentaho, siguiente los directorios Pentaho->server->pentaho-server. Esto nos permite ingresar a la dirección URL *localhost:8080/pentaho*, aquí debemos ingresar utilizando la cuenta de administrador o usuario establecida. Este modulo nos permite crear nuevos *Analysis Reports* o *Interactive Reports* en base a los cubos publicados previamente. En la Figura 5 de los Anexos se puede visualizar un ejemplo de reporte generado por Pentaho, utilizando los datos de Despachos, estos pueden ser exportados a formato PDF. Para ver otros reportes puede entrar en el repositorio del proyecto: *https://github.com/lAleRojasl/Proyecto2_BD2*, en la carpeta de *Reportes*.

# 5 Bitácora de trabajo

## 5.1 Alejandro Rojas

- 12-06-2017:
  - 1.5 horas - Descarga e instalación de Pentaho y Windows Server 2008 R2.
- 13-06-2017:
  - 4 horas - Diseño de base de datos y simulación. Diseño y simulación de productos y distribuidores.

- 14-06-2017:
  - 8 horas - Diseño y simulación de materiales, categorías, modelos, etc.
- 15-06-2017:
  - 5 horas - Simulación de ventas y despachos de productos.
- 17-06-2017:
  - 4 horas - Detalles de la simulación.
- 18-06-2017:
  - 3 horas - Conectar 2 máquinas virtuales con Windows Server 2008 R2.
  - 2 horas - Crear carpetas compartidas para archivos CSV, Excel y Replicaciones. Configuración correcta de la replicación de la base de datos de inventorio en la segunda máquina virtual.
  - 1.5 horas - Configuración correcta de la replicación de la base de datos de ventas en la primera máquina virtual.
  - 1 hora - Terminar llenado de la base de datos de inventarios.
- 19-06-2016:
  - 8 horas - Configuración de Pentaho y proceso de ETL / reportes.
- Total horas: 38 horas.

## 5.2 Saúl Zamora

- 12-06-2017:
  - 2 horas - Investigación sobre replicación de base de datos en SQL Server 2008 Enterprise.
- 13-06-2017:
  - 5 horas - Descarga de Windows Server 2008 R2 y configuración de la primera máquina virtual.
- 14-06-2017:
  - 1 hora - Configuración de la segunda máquina virtual con Windows Server 2008 R2.
  - 3 horas - Descarga de SQL Server Enterprise.
- 15-06-2017:
  - 2 horas - Instalación de SQL Server en ambas máquinas virtuales.

  - 2 horas - Configuración de carpetas compartidas en las máquinas virtuales para archivos CSV.

- 17-06-2017:

  - 5 horas - Intento de configuración de replicación de bases de datos con las máquinas virtuales de Windows Server y SQL Server.

- 19-06-2017:

  - 4 horas - Documentación.

- Total horas: 25 horas.

# 6 Comentarios finales

## 6.1 Estado final del programa

Se logró implementar gran parte del proyecto, entre estos puntos están:

- La simulación de datos de Ventas, Inventarios y Despachos de los 5 años establecidos, creando suficientes datos para alcanzar los mínimos establecidos en el enunciado del proyecto (6000 materiales, 100 dispositivos, 80000 movimientos, etc).

- Replicación funcional de las bases de ventas e inventarios, permitiendo reflejar automáticamente las modificaciones realizadas en ambas bases.

- Cubos y reportes mediante la herramienta Pentaho, se implementaron la mayoría de las consultas esperadas.

Las siguientes funcionalidades no se lograron implementar:

- Base de datos Post-venta: Se debía implementar una base de datos que guardara los datos recolectados por los productos IoT utilizados por los consumidores. Sin embargo con el objetivo de avanzar en las otras secciones del proyecto, se omitió la implementación de la simulación de estos datos. Consecuentemente las consultas relacionadas a este tampoco pudieron desarrollarse.

## 6.2 Problemas encontrados

- El principal problema encontrado fue en lo relacionado a comunicación entre las bases de datos, en donde al intentar crear la subcripción de una de las bases de datos, esta no encontraba a la otra. Para solucionar este problema se encontró que la principal causa suele ser el bloqueo de los puertos TCP 1434 y UDP 1433, por lo que se agregaron reglas de entrada a la configuración del Firewall. Adicionalmente, se corroboró que los servicios de *SQL Server Browser* estuvieran inicializados.

- Otro problema encontrado fue a la hora de importar los datos de los archivos CSV, la herramienta retornaba que no se permitian atributos nulos para ciertas columnas, sin embargo los archivos estaban generados de manera que no existiera ningún atributo nulo. Se descubrió que a la hora de generar dichos archivos, se estaba agregando un caracter de *newline* al final del archivo. Lo que se estaba tomando como una entrada nula.

# 7 Conclusiones y Recomendaciones

- Es importante tomar en cuenta la configuración en el firewall de Windows Server 2008 R2 a la hora de configurar máquinas virtuales con el propósito de que sean capaces de "verse entre sí".
- Revisar que los archivos CSV no posean filas en blanco al final del archivo. Puesto que a la hora de importar los datos, estas se toman como entradas nulas y el proceso falla.

# References

[1] Singh, S. (2017). SQL Server Performance Setting up Transactional Replication in SQL Server 2008 R2. [online] Sql-server-performance.com. Available at: http://www.sql-server-performance.com/2010/transactional-replication-2008-r2/

[2] VS (2017). Error Instalación SQL - Performance Counter Registry Hive Consistency. [online] Es.slideshare.net. Available at: https://es.slideshare.net/adictes/error-instalacin-sql-performance-counter-registry-hive-consistency [Accessed 18 Jun. 2017].

[3] YouTube. (2017). Replicar una Base de Datos de SQL Server 2008. [online] Available at: https://www.youtube.com/watch?v=ksnuYpE7A3s [Accessed 18 Jun. 2017].

[4] Pentaho Documentation. (2017). Connect to the Pentaho Repository from the PDI Client. [online] Available at: https://help.pentaho.com/Documentation/7.0/0H0/Connect_to_the_Pentaho_Repository_from_the_PDI_Client [Accessed 19 Jun. 2017].

[5] YouTube. (2017). Pentaho Analyzer Reports - Getting Started Analyzer. [online] Available at: https://www.youtube.com/watch?v=lZCBCXZ9BXI [Accessed 19 Jun. 2017].

# 8 Anexos

Figure 1: Diagrama de la base de datos de Inventario

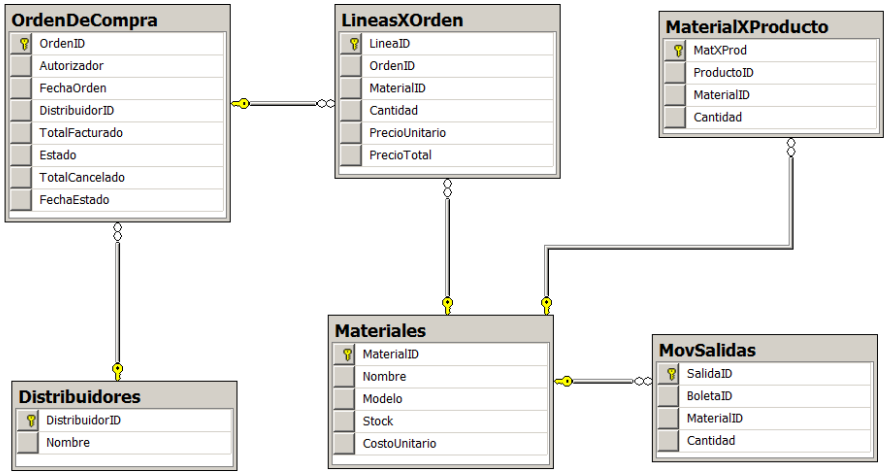


Figure 2: Diagrama de la base de datos de Ventas

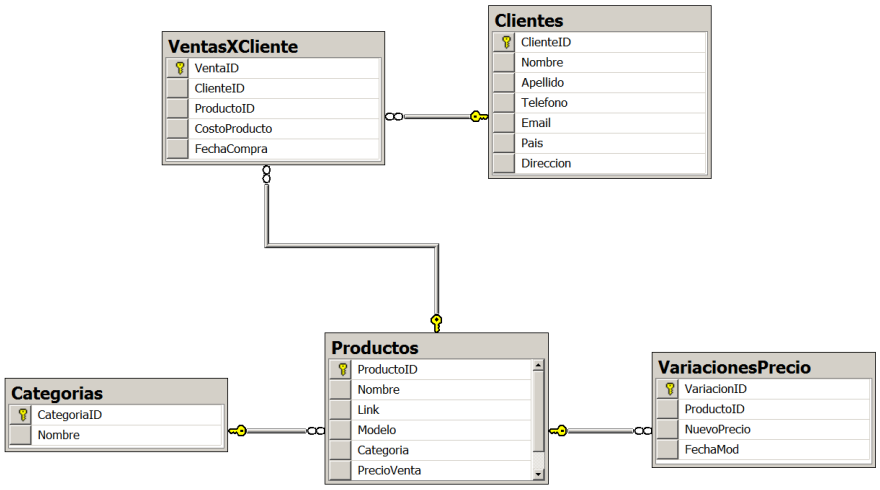

Figure 3: Configuración de conexión a Pentaho

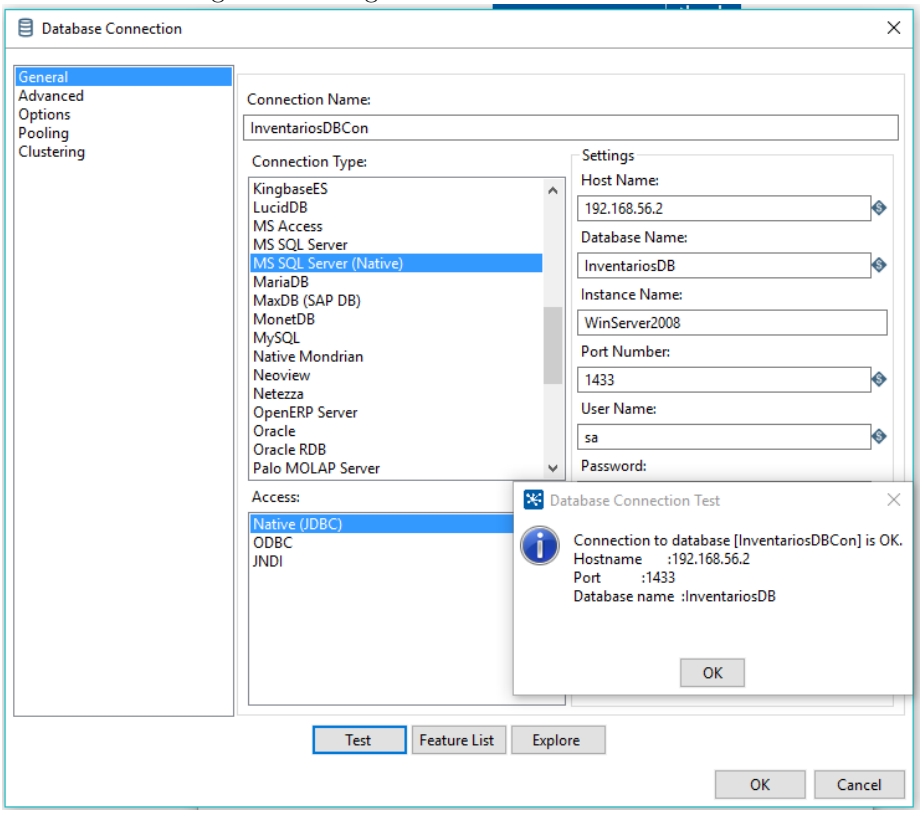

Figure 4: Estados de las bases de datos replicadas

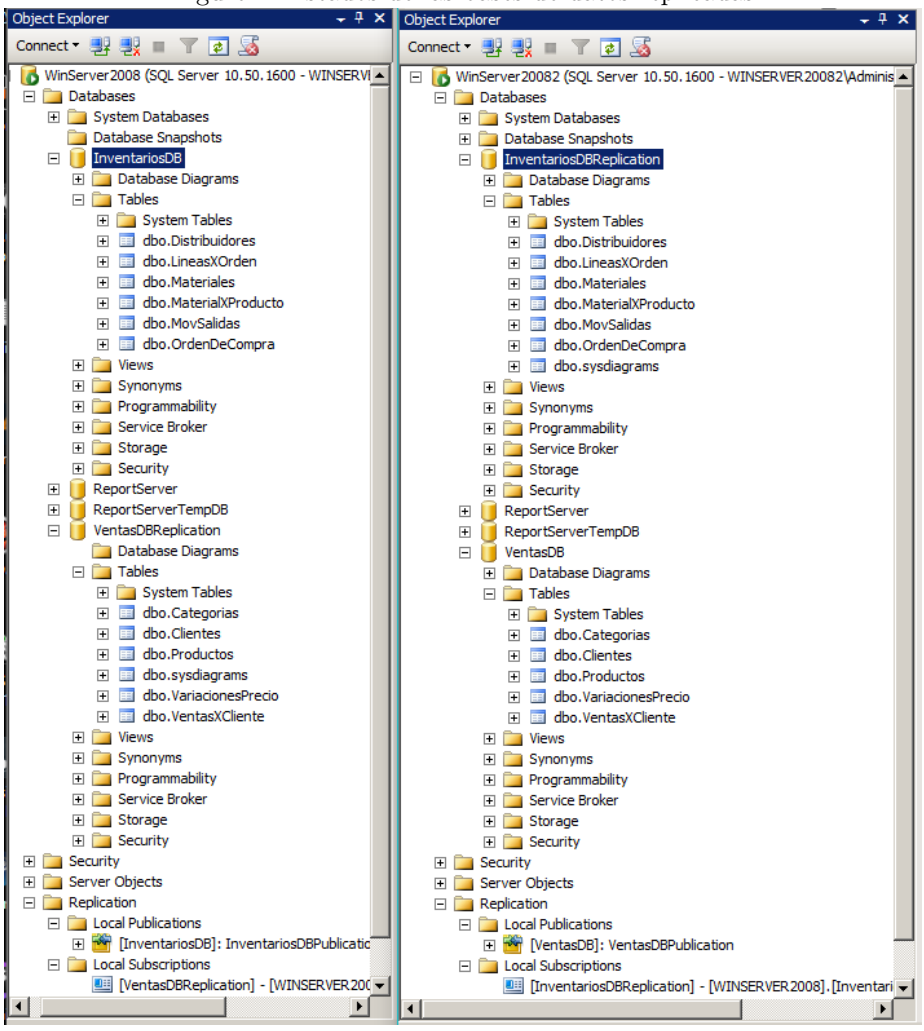


Figure 5: Reporte de Analisis de Despachos con dimensiones de Carriers y País

Analisis De Despachos

| | PaisDestino | | | |
|---|---|---|---|---|
| | Costa Rica | Guatemala | Nicaragua | Panama |
| Carrier | CantidadDespachada | CantidadDespachada | CantidadDespachada | CantidadDespachada |
| FedEx | 803 | 1.311 | 615 | 774 |
| Correos de Costa Rica | 729 | 1.303 | 622 | 765 |
| DHL | 767 | 1.241 | 595 | 809 |
| GoPato | 715 | 1.225 | 614 | 759 |
| UPS | 648 | 1.333 | 567 | 729 |