**Design Tip #174 ¿Tu organización necesita una caja analítica?**

Innumerables organizaciones han creado data warehouses dimensionales considerados como grandes éxitos dentro de sus organizaciones. Estos entornos de data warehouse respaldan los informes clave y los requisitos de análisis para la empresa. Muchos son capaces de dar apoyo al acceso autogestionado a los datos y a las capacidades de análisis para usuarios de negocio dispares.

Sin embargo, pese al éxito obtenido por estos data warehouses dimensionales, son a veces criticados por ser demasiado lentos en reaccionar a los nuevos requisitos, implementar nuevas fuentes de datos y respaldar las nuevas capacidades de análisis. A veces estas preocupaciones son exageradas ya que claramente se necesita un cierto período de tiempo para reaccionar a los nuevos requisitos, pero otras veces estas críticas son ciertas. Muchos data warehouses han crecido y evolucionado para llegar a ser entornos de misión crítica que respaldan los informes de empresa clave, cuadros de mando, tablas de puntuación y capacidades de acceso autogestionado a los datos. Debido a la naturaleza de misión crítica, el modelado dimensional, la dirección, el desarrollo de reglas ETL, y los cambios en los requisitos de gestión desembocan en una aprobación larga, ciclos de diseño y desarrollo para los nuevos requisitos y cambios. En muchos sentidos, estos desafíos son el precio del éxito.

El data warehouse es probable que esté muy estructurado, tremendamente diseñado, sujeto a reglas de negocio bien definidas, y estrechamente dirigido por la empresa. Gran parte de los datos del data warehouse se limpian y transforman de forma intensiva para asegurar que ofrecen una imagen fiel de lo que pasa actualmente en el negocio. Además, los datos del data warehouse se sincronizan frecuentemente con los entornos de producción vía cargas programadas regularmente. De esta manera, al final, es bastante rígido; simplemente lleva tiempo reaccionar a los nuevos datos y las solicitudes analíticas.

Todavía en el mundo competitivo de hoy en día, las organizaciones necesitan ser más ágiles. Quieren comprobar rápidamente las nuevas ideas, nuevas hipótesis, nuevas fuentes de datos y nuevas tecnologías. La creación de una caja analítica puede ser una respuesta apropiada a estos requerimientos. Una caja analítica complementa el data warehouse dimensional. No se destina a reemplazar el data warehouse, sino que se mantiene a su lado y proporciona un entorno que puede reaccionar más rápidamente a los nuevos requerimientos. La caja analítica no es un concepto realmente nuevo, pero en los recientes debates sobre grandes datos han puesto el concepto de nuevo en vanguardia. Normalmente se cree que una caja analítica es un área situada fuera de la infraestructura existente en el data warehouse o como un entorno separado adyacente al data warehouse. Proporciona el entorno y los recursos requeridos para apoyar las capacidades analíticas experimentales o de desarrollo. Es un lugar donde estas nuevas ideas, hipótesis, fuentes de datos y herramientas pueden ser utilizadas, probadas, evaluadas y exploradas. Mientras tanto, el data warehouse se mantiene como el pre-requisito de la fundación de datos que contiene los datos históricos precisos que los esfuerzos de la caja analítica hace girar alrededor y en contra.

A veces la clave de datos se alimenta del entorno de data warehouse existente en la caja analítica y se alinea con otros almacenes sin datos de almacenes de datos. Es un lugar donde las nuevas fuentes de datos pueden ser probadas para determinar su valor para la empresa. Un ejemplo de estas nuevas fuentes de datos puede ser la inteligencia de mercado adquirida externamente, atributos de los consumidores adquiridos externamente, o fuentes como interacciones en las redes sociales, interacciones con las aplicaciones móviles, dust móvil y la actividad web. A menudo sería muy oneroso traer estas nuevas fuentes de datos al entorno de data warehouse existente a menos que o hasta que se pruebe su validez. Los datos de la caja analítica no necesitan estar sincronizados sobre una base recurrente con un entorno de producción y estos conjuntos de datos expiran con el paso del tiempo.

Un objetivo clave de la caja analítica es comprobar cierta variedad de hipótesis sobre los datos y analíticas. De esta manera, no debería ser una gran sorpresa que la mayoría de proyectos de caja analítica terminen en fracaso. Eso significa que, la hipótesis no funciona como se esperaba. Esta es una de las grandes ventajas de la caja analítica. Los datos utilizados en estos fallos no tendrían que ejecutarse con el rigor esperado de los datos contenidos en el data warehouse. En este caso, el fallo es su propio éxito; cada fallo es un paso hacia la respuesta correcta.

La mayor parte de usuarios de negocio verán de manera acertada el data warehouse como la fuente de acceso a los datos de la empresa. Sus informes, cuadros de mandos/ tablas de puntuación y las solicitudes de auto-gestión ad hoc serán fácilmente respaldadas por el data warehouse. A los usuarios objetivo de la caja analítica se les llama normalmente “científicos de datos. Estos individuos son el pequeño grupo de usuarios de negocio suficientemente inteligentes para identificar las fuentes potenciales de datos, crear su propia “sombra” de base de datos, y construir análisis con objetivos especiales. A menudo estos individuos tienen que trabajar desconectados de la red. Ellos han elaborado y creado su propio entorno analítico en hojas de cálculo, conjuntos locales de datos, bajo los data marts del escritorio o donde quiera que el trabajo necesite ser llevado a cabo. La caja analítica reconoce que estos individuos tienen requerimientos reales. Les proporciona un entorno para trabajar en la red en un entorno dotado, respaldado, establecido, disponible, funcional y, de alguna manera más sutil, dirigido.

Disponer de las habilidades necesarias “in house” es esencial para el éxito de la caja analítica. Los usuarios de la caja analítica necesitas poder conectar con los datos con bastantes menos reglas que la mayoría de los usuarios de negocio. Son usuarios capaces de auto-proporcionarse datos tanto si vienen del data warehouse como si no. Son capaces de construir las analíticas y modelos directamente contra estos datos sin asistencia.

La caja analítica debería estar dirigida mínimamente. La idea es crear un entorno que vive sin la sobrecarga del entorno del data warehouse. No debería ser usado para respaldar las capacidades de la misión crítica de la organización. No debería ser utilizada directamente para controlar o respaldar ninguna de las capacidades centrales operacionales. Igualmente, no está destinada a ser utilizada para informes continuos o analíticas requeridas por los negocios de forma continua que respalden informes externos para cumplir las regulaciones financieras o gubernamentales.

Una característica importante de la caja analítica es que es de naturaleza transitoria. Los datos y el análisis vienen y van conforme son necesarios para respaldar los nuevos requerimientos analíticos. Los datos no persisten y no son regularmente actualizados vía las capacidades ETL. Los datos en la caja analítica normalmente tienen una fecha de expiración acordada. Así, cualquiera de los nuevos hallazgos o capacidades identificados como importantes para la organización y vitales para respaldar las contínuas capacidades necesitarán ser incorporadas en las operaciones de la empresa o en los entornos del data warehouse.

<http://www.kimballgroup.com/2015/05/design-tip-174-does-your-organization-need-an-analytic-sandbox/>