

Depósitos de datos y OLAP

Descripción del proyecto

El presente estudio procurará analizar las especificaciones y factibilidad de un proyecto que permita analizar el porcentaje de espectadores de las competiciones denominadas Copa Mundial de Fútbol por cada una de sus rondas a través de la historia. La solución que se propone para el almacenamiento y análisis de esta información es la instauración de un depósito de datos que permita a sus usuarios realizar consultas cuyas respuestas provean información confiable, oportuna y veraz sobre la cantidad – en porcentajes– de espectadores que una ronda de un mundial determinado obtuvo en una región acotada.

Análisis de viabilidad

Para proveer una resolución concreta y fiable respecto a la factibilidad y verosimilitud del desarrollo de este proyecto es preciso realizar un análisis sobre la posibilidad monetaria y técnica con que se dispone para la implementación de este. Además, es necesario estar consciente sobre las limitaciones y el alcance esperado del proyecto mismo, así como sus oportunidades y capacidades de crecimiento a lo largo del tiempo y maduración del sistema.

Presumiblemente, la información que el sistema necesita recolectar para su correcto funcionamiento atañe al porcentaje de la población de una región delimitada que sintonizó una ronda de un mundial específico. De este hecho se infiere que la máxima cantidad de registros que pueden ser almacenados sobre un solo mundial es de treinta –se prevé una duración de un mes para un mundial genérico–, aunado a la cantidad de información necesaria para representar todas las regiones concernientes –incluyendo en este conjunto el sur de Estados Unidos y el interior de la República mexicana– y la información necesaria para representar cada ronda que se lleva a cabo en un Mundial que es constante y, en principio, inmutable. Todos los datos que necesitan ser almacenados pueden ser descritos en formato de texto plano, es decir, no requieren recursos multimediales para su representación.

A partir de estas aseveraciones es posible determinar que, dada la relativamente poca información generada en lapsos extendidos donde en todo ese intervalo de tiempo la información permanecerá estática a menos que alguna actualización esporádica aparezca, un depósito de datos regular será capaz almacenar, administrar y manejar la información requerida para la implementación de este proyecto.

A este punto, se desconoce la capacidad adquisitiva de Infomex, mas se prevé que la inversión periódica que este proyecto requerirá para su desarrollo y mantenimiento oscila alrededor de los 300 dólares americanos al mes. Esta cantidad representa una inversión importante principalmente si la empresa no cuenta con el capital necesario para abastecer un desembolso continuo por este monto a menos que las utilidades generadas por esta inversión excedan esta cifra.

Para determinar si este proyecto es redituable es preciso indagar en los usos que el análisis de esta información o la información en sí misma podría desempeñar. Entre los principales es menester mencionar:

- Las televisoras que compran los derechos para transmitir en vivo las emisiones del Mundial representarán a los más potenciales clientes para este producto, el análisis de estos datos abrirá un amplio espectro mediante el cual las decisiones que éstas tomen sobre qué de qué partidos comprar derechos estarán mejor afianzadas y fundamentadas.
- Periódicos y revistas con secciones especializadas en deportes podrían utilizar información obtenida de este depósito de datos para crear redactar artículos basados en información veraz pero sobre todo obtenida de manera oportuna, lo cual resulta importante publicaciones de estilo diario.
- Sitios interactivos para entusiastas del deporte podrían basar su contenido en información obtenida de este depósito para proveer servicios propios de análisis, pronosticación de eventos futuros, ajustes de puntos de apuestas, entre otros servicios.
- Empresas en general interesadas en publicitarse podrían hacer uso del análisis para identificar patrones de tendencias que les permitan tomar una decisión de cuándo y dónde aumentar su propaganda.

Los posibles ámbitos en que este proyecto ofrece servicios importantes son tan variados como potencialmente fructíferos, razón por la que se concluye la redituabilidad de este servicio. No obstante, es solemne identificar, en adición, las limitaciones que esta solución mantiene; la primera de ellas es que su alcance es limitado, con respecto a que únicamente se guardará información sobre la zona sur de los Estados Unidos de América y México, por lo tanto, se considera que es un proyecto presumiblemente nacional con pocas implicaciones globales por el momento. Además, el análisis de los *ratings* se limita únicamente a resolver consultas referentes a las rondas de cada mundial, es decir, no ofrece información más específica que podría ser importante como las selecciones que juegan cada una de las rondas o el país en que se festeja el mundial, datos que podrían ser de fundamental importancia en muchas decisiones

Objetivos e hipótesis

El presente proyecto pretende proveer una solución a la necesidad que Infomex plantea referente a analizar el porcentaje de personas de una población que ven la Copa Mundial de Fútbol, para este fin la solución propuesta es modelar la información en un depósito de datos de tal manera que sea sencillo, plausible y eficaz integrar datos históricos con que se cuentan de mundiales pasados, registrar datos de nuevos mundiales y realizar consultas oportunas que favorezcan la toma de decisiones empresariales que se basen en información de este tipo.

Se presume que con la instauración de este sistema se incrementará la eficiencia del análisis de los datos en este sector y, consecuentemente, la toma de decisiones; la razón de esta asunción radica en el hecho de que al mantener toda la información en un sitio que engloba y es capaz de manejar información variada sin importar su fuente, los empleados podrán concentrar su energía únicamente en interpretar los resultados y plantear soluciones coherentes con base en ellas en lugar de invertir tiempo y energía tratando de analizar manualmente las múltiples fuentes a mano.

Especificaciones técnicas

El depósito de datos a implementar cuenta con las siguientes características:

- El proceso, es decir, la actividad de la cual se pretende extraer información, elegido corresponde al llamado *rating*, que representa el porcentaje de personas que sintonizan por cualquier medio el Mundial de Fútbol.
- El gránulo, que corresponde a las unidades más pequeñas que se almacenarán dentro del depósito de datos y que representarán la información sobre la actividad a modelar, decidido responde a la cuestión de solicitud de Infomex: se desea almacenar información sobre las estadísticas de cada **ronda** de cada **mundial** en cada **región** de México y Estados Unidos (Sur)
- Las dimensiones sobre las cuales se caracterizará la actividad elegida son tres:
 - Dimensión Ronda: representa cada una de las fases que pueden ser jugadas en un Mundial y están etiquetadas como *Eliminatorias*, *Octavos de final*, *Cuartos de final*, *Semifinal* y *Final*.
 - Dimensión Región: describirá la información geográfica básica que represente a las zonas aludidas previamente con cierta información de interés para este tema como la cantidad total de habitantes con las que cuenta. Puede ser categorizado en la siguiente jerarquía: región → estado → país.
 - Dimensión Mundiales: funciona de manera equivalente a como lo haría una dimensión tiempo, almacena información temporal respecto a cuándo ocurrieron sus hechos, pero más importante, almacena datos no calculables, por ejemplo, el día de la semana en que tuvo lugar o si coincidió con algún día feriado. No es redundante porque únicamente puede jugarse un Mundial por año, así, si una fecha buscada señala al año 2014 se puede inferir que se refiere al mundial de Brasil. Puede ser jerarquizada de la siguiente manera: día → semana → año
- Finalmente, la información que se pretende almacenar para medir las estadísticas y sobre la cual se realizarán los análisis pertinentes es el *rating*, una medida que describe el porcentaje total de habitantes de una región que vio los partidos de una ronda de algún Mundial.

A continuación, se ofrece una descripción gráfica de cómo las dimensiones enumeradas en el punto anterior se encuentran organizadas de acuerdo a sus jerarquías naturales, donde el elemento más a la izquierda representará el grado menor con que cada dimensión será almacenada en el depósito de datos.

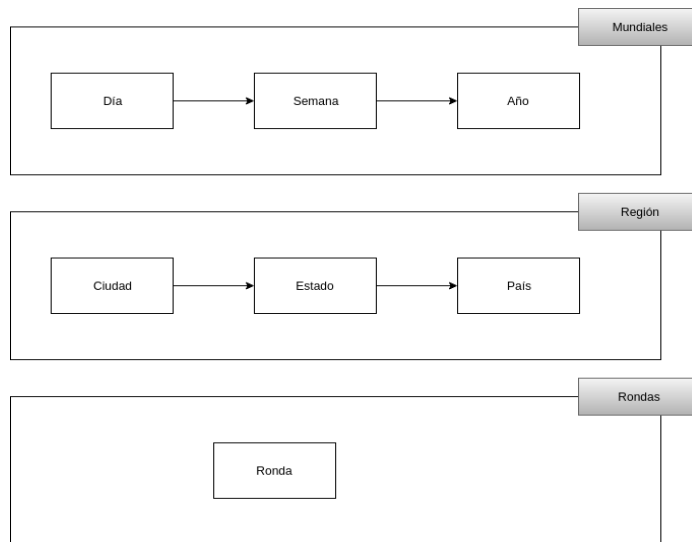


Figure 1. Esquema de dimensiones

En seguida, la Figura 2 revela una representación gráfica de la relación que mantienen estas dimensiones con la tabla de hechos (al centro) cuyas tuplas serán la razón central de la existencia del depósito de datos, en la que, además, se almacena la medida con que se analizará cada uno de estos hechos. Esta relación en forma de estrella describe a su vez las tablas de dimensión con sus debidas jerarquías expresadas previamente en conjunto con atributos auxiliares que serán de utilidad en el análisis.

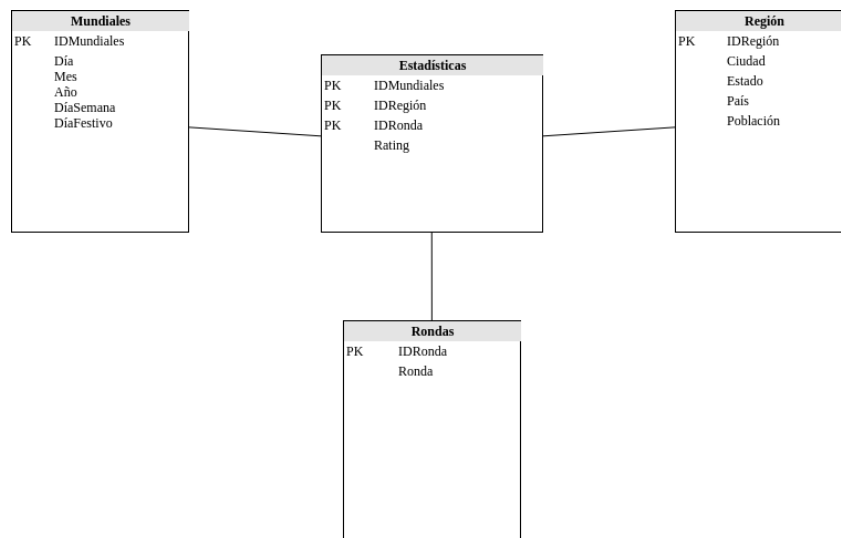


Figure 2: Esquema multidimensional en estrella

A partir de estos esquemas es factible entender la estructura propuesta para el diseño del depósito de datos, sin embargo, resulta fundamental también demostrar cómo es que estos datos serán utilizados en conjunto para proveer información útil para la toma de decisiones. Debido a que cada registro de la tabla de hechos es obtenido a partir de la relación de los datos encontrados en tres dimensiones se dice que cada nivel de agregación de este conjunto de dimensiones es un cubo.

A continuación, la Figura 3 expresa gráficamente una instancia, como ejemplo, de uno de estos cubos donde, particularmente se parametrizan las tres dimensiones por su gránulo más fino, en este

caso, de Región se obtiene la ciudad Puebla, de Rondas se obtiene la ronda equivalente a las semifinales que se llevan a cabo en un día específico sobre la dimensión mundiales. Dentro del cubo concéntrico se puede apreciar una cantidad en formato de porcentaje, éste representa el *rating* que este análisis persigue.

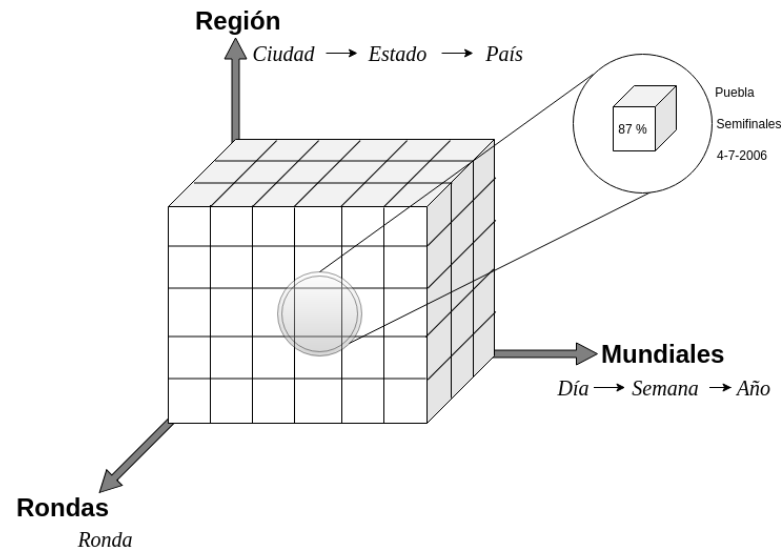


Figure 3: Instancia del cubo

En resumen, este pequeño cubo es obtenido como respuesta a la consulta “Rating que consiguieron las semifinales del 4 de julio de 2006 (Italia vs. Alemania en el mundial de Alemania 2006) en la ciudad de Puebla”, cuyo resultado será obtenido inmediatamente mostrando que el 87 % de la población poblana sintonizó ese partido.

Estas consultas también pueden ser expresadas en lenguaje SQL mediante las cuales es posible seleccionar qué tan específica la búsqueda debe realizarse, qué restricciones deben ser consideradas y la forma en que estas deben ser desplegadas. En los siguientes párrafos se ilustran algunos ejemplos de consultas que pueden llevarse a cabo y su respectiva equivalencia en SQL.

Proporcione el porcentaje promedio de espectadores que vieron el mundial de 2014 en cada país

```
SELECT Región.País AS País, AVG(Estadísticas.Rating) AS Rating
FROM Estadísticas, Región, Mundiales m
WHERE Estadísticas.IDMundiales = m.IDMundiales
      AND Estadísticas.IDRegión = Región.IDRegión
      AND m.Año = 2014
GROUP BY Región.País
```

Encuentre el porcentaje de personas de Veracruz que vieron la final del mundial de Sudáfrica el 11 de julio de 2010

```
SELECT Rating.Rating
FROM Estadísticas Rating, Mundiales m, Región, Rondas ron
WHERE Estadísticas.IDMundiales = m.IDMundiales
      AND Estadísticas.IDRegión = Región.IDRegión
      AND Estadísticas.IDRonda = ron.IDRondas
      AND m.día = 11-07-2010
      AND reg.Estado = 'Veracruz'
      AND ron.Ronda = 'Final'
```

Provea el año del mundial que mejor promedio de rating ha obtenido en Estados Unidos

```
SELECT AverageRatings.Mundial, AverageRatings.PromedioEspectadores
FROM (m.Año AS Mundial, AVG(Estadísticas.Rating) AS PromedioEspectadores
FROM Estadísticas, Region, Mundiales m
WHERE Estadísticas.IDMundiales = m.IDMundiales
AND Estadísticas.IDRegión = Región.IDRegión
AND Región.Pais = 'Estados Unidos'
GROUP BY m.Año) AS AverageRatings
WHERE AverageRatings.PromedioEspectadores = (
    SELECT MAX(Promedio)
    FROM (
        SELECT AVG(Estadísticas.Rating) AS Promedio
        FROM Estadísticas, Region, Mundiales m
        WHERE Estadísticas.IDMundiales = m.IDMundiales
        AND Estadísticas.IDRegión = Región.IDRegión
        AND Región.Pais = 'Estados Unidos'
        GROUP BY m.Año
    )
)
```

Metodología para el desarrollo del proyecto

Para la implementación del diseño actual se compararon las principales características de dos de las herramientas más populares actualmente para la implementación de depósitos de datos, Oracle 12c y Panoply.

Oracle ofrece una solución estable y ampliamente probada y utilizada a lo largo de los años, entre sus ventajas se encuentran sus capacidades de simplificación de consultas complejas, estructuración y agendación de paquetes de código que permiten realizar el mantenimiento de manera oportuna, entre muchas otras ventajas. Panoply por su parte es una alternativa relativamente nueva pero que se ha ganado su lugar en el mercado a pulso, su fuerte más grande está en la extrema facilidad de integración que ofrece para una extensa variedad de fuentes de las que presume no necesitar conocimiento humano para integrar; además, utiliza técnicas de aprendizaje de máquina y procesamiento de lenguaje natural para aprender y realizar adecuaciones pertinentes sobre el modelo, lo cual otorga un gran potencial de mejora a lo largo de los años.

Ambas opciones proveen servicios destacables y más que suficientes para el alcance que esta aplicación requiere, sin embargo, el precio que ambas exigen varía. Panoply ofrece un plan de 350 dólares al mes por el mantenimiento y consultas del depósito de datos; mientras que Oracle no ofrece un sistema de preciado tan directo, sino que se basa en una extenuante cantidad de factores específicos para cada depósito de datos, conseguir un precio preciso y justo para los requerimientos específicos necesita de una certificación en fijación de precios Oracle, pero para el alcance de esta aplicación se estiman los mínimos requerimientos posibles que resultan en un precio inicial superior a los 500 dólares más un cargo mensual de 388 dólares.



Autonomous Data Warehouse Cloud - BYOL	\$582	\$388	
>  Autonomous Data Warehouse	\$222	\$148	
>  Autonomous Data Warehouse - BYOL	\$360	\$240	

Figure 4: Precios Oracle

De esto se concluye que la mejor alternativa será optar por Panoply con programación en lenguaje Java integrando cualquier herramienta de análisis de datos que se utilice ya en Infomex.

Conclusiones

La implementación del proyecto es factible y puede ser llevada a cabo en un amplio espectro de posibilidades de entre las cuales se ha elegido la instauración de un depósito de datos instalado en Panoply modelado en secciones anteriores debido principalmente a su facultad natural de análisis de hechos basados en dimensiones. Podría argumentarse que dada la relativamente poca cantidad de información a almacenar y los espaciados intervalos entre actualizaciones relevantes un depósito de datos resulta demasiado para representar este problema y que una base de datos tradicional sería suficiente, sin embargo, para que una base de datos ofrezca un nivel de análisis equiparable al de un depósito de datos se requiere un trabajo humano y computacional que superaría la inversión de un depósito de datos.

A pesar de que el modelado descrito en el presente estudio cubre las partes principales y el objetivo central del proyecto es menester reconocer que aún cuenta con algunas limitaciones que fueron descritas en la sección *Análisis de viabilidad*, no obstante, el conocimiento de las mismas no restringe el desarrollo, sino que propone oportunidades de crecimiento del proyecto, de tal suerte que una vez implementada y probada –etapa de *organización y dirección* de un proceso administrativo– al llegar a la etapa de *control*, si las hipótesis formuladas en esta *planeación* resultan ser comprobadas entonces se pueden realizar ajustes que permitan recolectar, almacenar y analizar datos más específicos según sea requerido y expandir las regiones soportadas para una visión más globalizada.

Referencias y bibliografía

- Capterra.com. (2019). Oracle Database vs Panoply - 2019 Feature and Pricing Comparison. [online] Available at: <https://www.capterra.com/data-warehouse-software/compare/5938-168034/Oracle-Database-vs-Panoply> [Accessed 18 May 2019].
- Oracle Technology Global Price List. (2019). [ebook] Available at: <http://www.oracle.com/us/corporate/pricing/technology-price-list-070617.pdf> [Accessed 18 May 2019].
- Panoply. (2019). Data Warehouse Automation Done Right. [online] Available at: <https://panoply.io/platform/> [Accessed 18 May 2019].
- Softwaretestinghelp.com. (2019). Top 10 Popular Data Warehouse Tools and Testing Technologies. [online] Available at: <https://www.softwaretestinghelp.com/data-warehouse-tools/> [Accessed 17 May 2019].