Pontificia Universidad Católica Madre y Maestra
PUCMM

Asignación
Proyecto Final

Asignatura:
Base de Datos II

Presentado por:
Richard García
Michael Romano
Michael Gaillard

Presentado a:
Máximo E. Pérez M.

Santiago, República Dominicana.

INTRODUCCION

El trabajo trata sobre el juego de béisbol. El béisbol se juega en un campo cubierto por grama natural o artificial. Excepto la línea de corredor que es la línea donde los jugadores corren para alcanzar las bases. Hay cuatro de estas bases. Estas bases forman un cuadrado y están en los extremos. El área se le llama diamante.

El juego consiste en golpear una pelota con un bate y correr tratando de alcanzar darle la vuelta al diamante y volver a la posición inicial para anotar puntos. Hay nueve innings y el equipo que anote más carreras es el ganador. Existen los infielders que son la defensiva que se encuentra dentro del diamante y los outfielders que son la defensiva que se encuentra afuera del diamante.

El sistema de base de datos de nosotros permite hacer diferentes análisis sobre los jugadores y equipos de la liga Americana y Nacional (y cualquier otra liga que se agregue, gracias a la flexibilidad del diseño). Una gran parte de esta información la sacamos en baseball-reference.com, una página que es rica en contenido de béisbol y la cual fue de mucha ayuda para la realización de nuestro proyecto. No fue difícil crear un programa donde le podemos pasar todos los datos de los diferentes equipos y jugadores para insertar estos datos en nuestra base de datos. Esta parte es importante hacer si en el futuro se desea poder actualizar los datos que están contenido en la bases de datos sin tener que hacer demasiado trabajo. Lo ideal sería tenerlo en formato XML o JSON y actualizarlo.

DESCRIPCION GENERAL DEL PROYECTO

Muchas cadenas de deportes y equipos necesitan una forma de analizar la información de todos los jugadores de cada temporada para luego poder tomar decisiones en base a estos análisis que generen. Para lograr dicho reto se implementaran los conocimientos adquiridos en la materia de Base de Datos II para diseñar e implementar un ambiente de análisis de datos que permita eficientizar el proceso de análisis y la toma de decisiones nivel de bate del equipo, picheó del equipo, y fildeo del equipo por temporada (año).

PLAN/ESTRATEGIAS DEL PROYECTO

Para cumplir con nuestra meta se llevaron a cabo los siguientes pasos:

1. Diseño e implementación del ambiente operacional (realizado en PostgreSQL): Tomando en cuenta los procesos que se iban a modelar, se diseñó el modelo operacional respectivo de tal modo que cumpla con las normas de una base de datos relacional.

2. Inserción de datos a nivel operacional: Una vez creado e implementado el ambiente operacional de nuestro proyecto, se procedió a popular dicho ambiente con la inserción de datos en sus tablas correspondientes, para luego ser utilizados como base en el ambiente de data warehousing. Para este paso obtuvimos información verídica gracias a la página baseball-reference.com y para automatizar el proceso de inserción de datos se crearon varios scripts en Python para insertar los datos en las tablas del ambiente operacional.

3. Diseño e implementación de un ambiente de data warehousing (realizado en MySQL): Teniendo el ambiente operacional funcionando ya, se analizaron los procesos que se iban a modelar y se diseñó el ambiente dimensional. En nuestro caso son: bateo del equipo, picheo del equipo, y fildeo del equipo.

4. Proceso ETL: esta fue la parte más interesante debido a como nosotros lo hicimos. Ya teniendo el ambiente operacional en funcionamiento y el diseño dimensional, se realiza el proceso de ETL (Extraemos la data del ambiente operacional, Transformamos dicha data de modo que se pueda insertar en el ambiente dimensional, Load se carga en el ambiente dimensional junto a todos los otros datos previamente montados). Para nuestro proyecto dicho proceso fue realizado con la ayuda del lenguaje de programación Python, muy interesante!

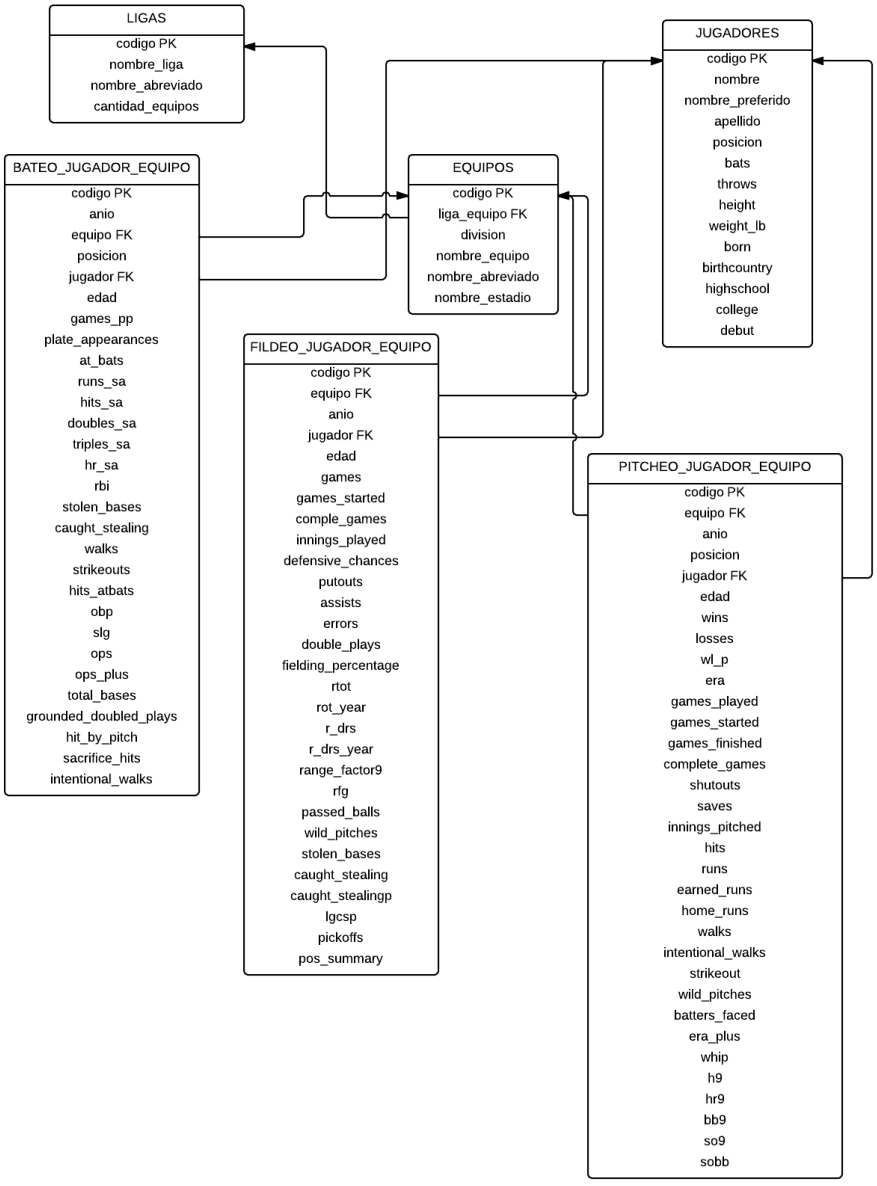
5. Análisis de datos: Ya con ambos ambientes, operacional y dimensional en funcionamiento, procedemos a realizar los análisis en base a los procesos modelados. Para llevar a cabo dicha tarea se utilizó la herramienta Pentaho Report Designer.

MODELOS SELECCIONADOS

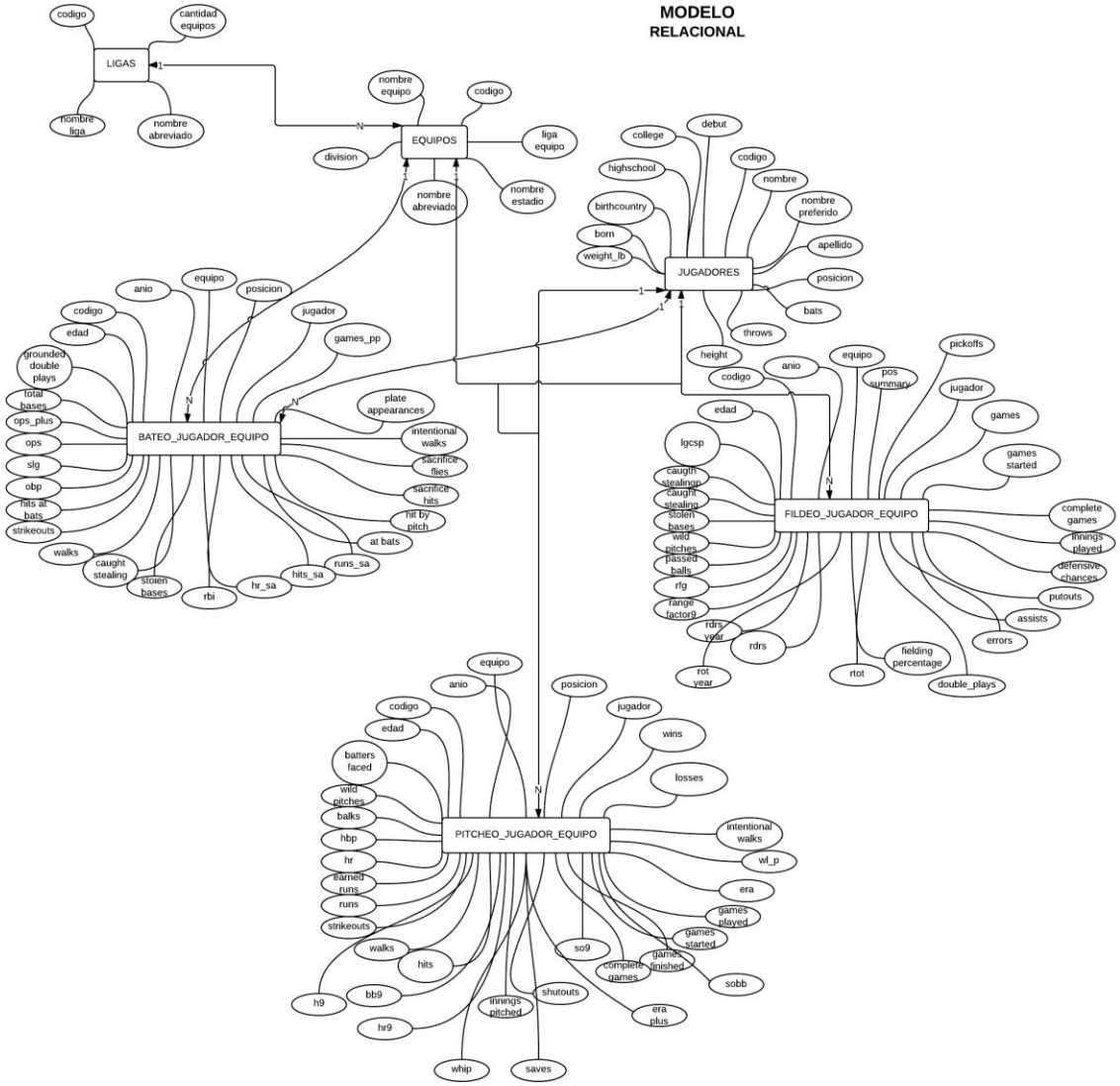
El modelo que se utilizó para la construcción del ambiente dimensional fue el modelo estrella. Básicamente este modelo está compuesto por varias tablas de hechos de las cuales hacen referencia a dimensiones comunes.

Cada tabla de hecho representa un proceso a modelar de nuestro sistema. Obviamente como nuestro proyecto trata sobre béisbol, las medidas más útiles para incluir en una tabla de hechos son las medidas aditivas que son aquellas medidas que pueden ser sumadas como por ejemplo carreras impulsadas, hits de un jugador, entre otras. En diferencia una tabla de hecho, una tabla de dimensión está compuesta por un conjunto de atributos descriptivos, que nos ayudan a realizar ese estudio proceso aportando información sobre los atributos de la tabla de hechos. A continuación se muestran los diseños de los modelos que también están incluidos en el archivo rar por si se desean observar con más detalle.

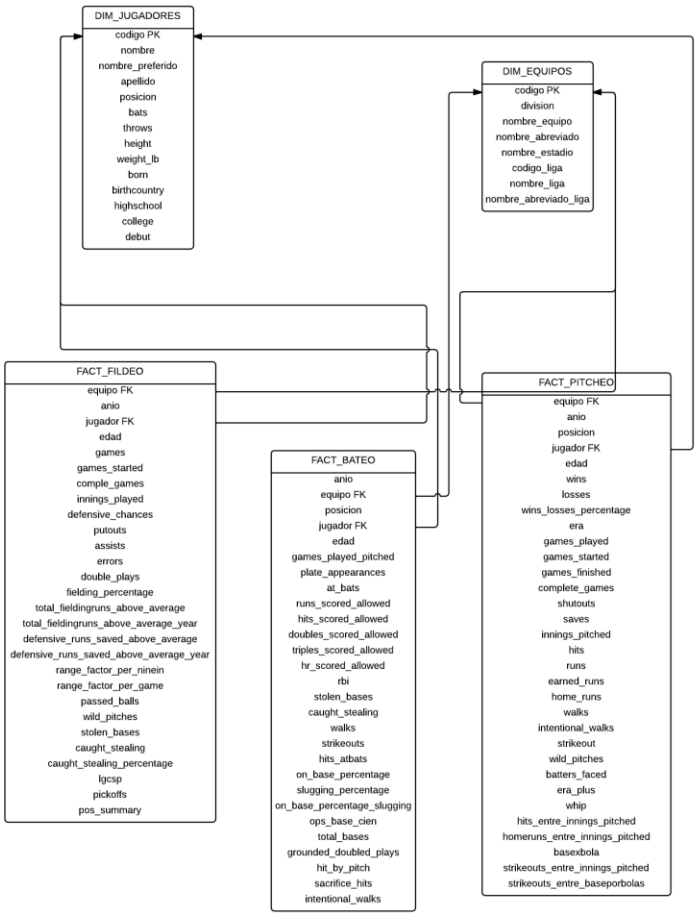
Modelo de Datos Operacional



MODELO RELACIONAL



Modelo Dimensionales



DESCRIPCION DE LOS OBJETIVOS DEL PROYECTO

Este proyecto tiene como objetivo aplicar para el área de major league baseball el uso de inteligencia de negocio mediante la creación de un ambiente operacional y dimensional (datawarehousing). Para la generación de los reportes se utilizó Pentaho Report Designer. Al final de dicho proyecto los usuarios podrán analizar data de una temporada y tomar decisiones en base a lo analizado.

PROCESOS MODELADOS

Los procesos a modelar son los siguientes: team batting, team pitching, team fielding. Los tres procesos tienen un nivel de grano alto pero a la vez sumamente detallado. En cuanto a stats se refiere a todos los atributos proporcionados por la página que nos sirvió de ayuda (baseball-reference.com)

Team Batting (Bateo del equipo): Se tomaran los stats de cada jugador de cada equipo.

Team Pitching (Picheo del equipo): Se tomaran los stats de cada pitcher de cada equipo.

Team Fielding (Fildeo del equipo): Se tomaran los stats de fildeo de cada jugador de cada equipo.

DESCRIPCION DE LAS TABLAS DEL AMBIENTE OPERACIONAL

NOMBRE TABLA	**DESCRIPCION**
Ligas	Tabla encargada de guardar los datos pertenecientes a las ligas del béisbol.
Equipos	Tabla encargada de guardar los datos pertenecientes a los de equipos que existen en las grandes ligas.
Jugadores	Tabla encargada de guardar los datos pertenecientes a los jugadores de las grandes ligas.
Bateo_jugador_equipo	Tabla encargada de guardar los datos pertenecientes a los stats de batting de cada jugador por temporada que existen en las grandes ligas.
Pitcheo_jugador_equipo	Tabla encargada de guardar los datos pertenecientes a los stats de picheo de cada pitcher por temporada que existen en las grandes ligas.
Fildeo_jugador_equipo	Tabla encargada de guardar los datos pertenecientes a los stats de fildeo de cada jugador por temporada que existen en las grandes ligas.

DICCIONARIO DE DATOS AMBIENTE DWH

Bateo_Jugador_Equipo	
games_pp	Esta columna guarda la cantidad de juegos jugados por el jugador ya sea como pitcher o no.

plate_apearrances	Este valor es la cantidad de veces que el jugador ha estado en el plato incluyendo base por bolas, las veces que ha sido golpeado por el pitcher y base por bola intencional.
at_bats	Cantidad de veces que ha estado al bate.
Rbi	Cantidad de carreras empujadas.
Walks	Base por bola.
hits_attacks	Porcentaje de hits por cantidad de veces al bate.
Obp	Porcentaje de embasado.
Slg	Index que se calcula dando un valor a cada diferente base lograda entre la cantidad de turnos al bat.
Ops	Suma de porcentaje de embasado mas index por base.
Hbp	Veces golpeado por un pitcher.
grounded_dp	Doble plays.

Pitcheo_Jugador_Equipo	
wl_p	Porcentaje de ganada y perdidas
Era	Carreras permitidas entre cantidad de innings
Shotouts	Juegos completos sin permitir carrera
Hpb	Cantidad de veces que golpea al bateador.
era_plus	Un ajuste de era (100*(lg(era)/era))
H9	Hits por ining.
Hr9	Home runs por ining.
Bb9	Base por bolas por ining.
So9	Ponches por inning
Sobb	Cantidad de ponches entre base por bolas

Fildeo_Jugador_Equipo	
defensive_chances	Oportunidades de hacer atrapar la bola. (Errores + asistencias mas putouts)
Rtot	Cantidad de carreras arriba o abajo del promedio que el fielder logro por por jugada.
r_drs	
range_factor9	Putouts y asistencias por ining
Rag	Putouts y asistencias por juego
caught_stealingp	Porcentage de robos de base frustrados
Lgcps	Porcentaje de frustracion de robos de base por liga.
Pos_summary	Posiciones jugadas

DESCRIPCION DEL ETL

El ETL lo creamos utilizando Python como lenguaje de programación. En esta aplicación nos conectamos a ambas bases de datos la del ambiente operacional y la del ambiente dimensional.

Cada vez que queremos popular las tablas simplemente ejecutamos la aplicación (que básicamente es una script) la cual se encarga de seleccionar datos de las distintas tablas del ambiente operacional e insertar estos datos en las dimensiones del ambiente dimensional tomando en cuenta que no se inserten datos que ya estaban previamente insertados.

CREACCION DE LAS TABLAS DEL AMBIENTE OPERACIONAL Y DIMENSIONAL

Al tener demasiada información, decidimos no ponerlas aquí ya que el archivo sería bastante largo. Por ende los créate y inserts están también contenidos en el rar en las carpetas “Queries” y “Query Generators”. También los Scripts del ambiente operacional y el dimensional se encuentran en la carpeta “Scripts”.

ANALISIS

Aquí tenemos la descripción de los 30 análisis, ahora bien, debido a que serían demasiados print screens, los resultados de los análisis los he puesto en una carpeta llamada “Análisis”. Están enumerados en esa misma carpeta por orden.

Team Batting

1. El bateador que tiene la mayor cantidad de homeruns por equipo.

2. Bateador designado que tenga cantidad de carreras y carreras empujadas por equipo, si es que existe un bateador designado en dicho equipo.

3. Los bateadores con el mayor promedio en cada equipo y que esos equipos sean ordenados en forma descendente.

4. La edad media de los bateadores de la liga completa por año.

5. La edad en que los jugadores tienen el mejor desempeño de bateo.

6. El jugador que más se ha ponchado por año.

7. El jugador que se ha robado más bases por año de cada equipo.

8. El jugador que tiene mejor rendimiento robándose una base.

9. El jugador más temido de la liga por año. (Bases por Bolas intencionales)

10. El porcentaje de que un jugador cuando batee haga carreras, de cada jugador por cada equipo de cada año.

Team Pitching

11. EL pitcher que ha ponchado más jugadores de cada equipo por año (De mayor a menor).

12. El pitcher con el promedio de innings pichados por juego de cada equipo por año (de mayor a menor).

13. El pitcher que ha abierto más partidos de cada equipo, por año.

14. El abridor que tiene la mejor era de cada año.

15. El cerrador que ha salvado más juegos por año.

16. El pitcher que más juegos ha ganado por año.

17. El equipo que tiene el mayor promedio de porcentaje de ganadas y pérdidas de sus pitchers. Esto da un indicio a la probabilidad de un equipo ganar un juego.

18. El peor pitcher del año. Tomando en cuenta los atributos "negativos" dígase errores. (wild_pitches + hpb + balks + runs + losses).

19. La liga que tiene los pitchers que han ganado más juegos por año.

20. El pitcher que ha mas bateadores se ha enfrentado de cada equipo por año.

Team Fielding

21. El jugador que ha cometido más errores de cada equipo por año.

22. EL jugador que ha cogido más outs y su posición de cada equipo por año.

23. El fielder que más asistencias tenga por año.

24. El equipo que tenga más doble plays por anio.

25. El jugador con el mejor fielding percentage de cada año.

26. El equipo que más ha atrapado a jugadores que tratan de robarse una base.

27. La cantidad de putouts por juego de cada jugador de cada equipo por año.

28. El jugador que tenga el promedio más alto de atrapar a un corredor.

29. El equipo que tiene el mejor fielding percentage de cada año.

30. El jugador que ha tenido más defensive chances por año.

NOTA: si desea ver los select de cada análisis lo puede encontrar dentro del rar en un archivo llamado "BDDII PF.txt". También si desea ver los proyectos generados por Pentaho están en la carpeta "Pentaho".

CONCLUSION

Para concluir este reporte podemos decir que para lograr una buena arquitectura de Data warehouse es necesario que en cada paso del proceso del diseño del mismo se tomen las decisiones correctas, desde la elección del nivel de grano que se utilizara basándonos en nuestras necesidades finales de análisis hasta los manejadores que se utilizaran.

En nuestro caso decidimos utilizar MySql y PostgreSQL como manejadores ya que nos facilitaba el curso del proyecto dadas las herramientas que teníamos a mano ya como desarrolladores así como la producción de un ETL con mayor comodidad gracias a conocimientos previos del lenguaje que se utilizó para la realización del mismo.

BIBLIOGRAFIA

Wikipedia: http://es.wikipedia.org/wiki/B%C3%A9isbol

Baseball-reference: www.baseball-reference.com