Project MoneyBall

Technical Design Document

Submitted by:

Nikhil Raikar
Cheryl Jacob
Vamshi Bhushanaboina
Pooja Ajit

Under the guidance of:
Stefan Willi Hart

Table of Contents

1. Document Information

Versioning: v0.8 – ready for review; v0.9 – ready for sign-off; 1.0 sign-off

Version No.	Date	Author	Comment	Reviewer	Release
v0.8	04/03/2018				

Table 1: Version History

Ref.	Document	Notes/Link
1	SAP TDD Template	
2	Lehman's Database V2016	http://seanlahman.com/baseball-archive/statistics

Table 2: References

Abbreviation	Description
Association	Relation between data

Table 3: Abbreviations and Special Terms

No.	Assumption	Referred Section	Logged by	Logging Date
1	Batting is more important than fielding and pitching			
2				
3				

Table 4: Assumptions

No.	Open point	Referred Section	Owner	Expected update date
1	More analysis can be done			
2	More features can be added to make the analysis better.			
3				

Table 5: Open Points

2. Overview

This document is a generic Technical Design Document for use. It provides guidance and template material which is intended to assist the relevant management or technical staff about the implementation procedures taken up in this project. This document clearly explains all of the technical design implementations, the operations performed on the dataset and the analysis results obtained from the data.

In the below sections this design document takes you through clearly as how the analysis were formed, how our data model was created and what are all the design changes and the implementations done with the data.
Lehmann's BaseBall data is the dataset used in the implementation. You can also fing the reference to the dataset in[2].
In section 2 we clearly explain as how the data is loaded from the user's local database into the SAP HANA database using various forms of process logic.
In section 3 we take you through our technical interface. The interface being the core of the application, all of the analysis logic, various forms of design implementations have been clearly mentioned in this section.
In section 4 we take you though our data model, you can find information's about the logical and physical data models and their uses.
In section 5 and 6 we explain briefly the functional and non functional requirements. Various topics such as performance, availability, maintainability are discussed.
Security being one of the main concerns of any application we take you through some of these measure in section 7.

2.1. Service Scope

2.2. Technical Description of the Application

The applications User Interface is built using basic Javascript, HTML and css languages, but the core of the application, the analysis part is designed with the help of predictive analysis tools. The data which is passed to the predictive analysis is modelled using various features and association properties found within the dataset.

2.3. Processing Logic

Table 4 Processing Logic

Layer	Processing Logic
Calculation Logic	• The different column views such us Batting, Fielding, Pitching and Salary data are added into the SAP HANA catalog. The necessary SQL queries are implemented to sort and derive only the data that is needed for our analysis. The data modelling process is implemented by creating calculation views based on the catalog data to refine and clean the data to perfectly suit our needs. In this project we have used CUBE type calculation view with star join to merge all the different column views on the database. • The data is then loaded into the SAP predictive analysis tool to perform analysis on the baseball data. After the calculation, the data is populated back to the catalog and processed further. • For example, In Predictive analysis using ABC clustering, we have used calculation views to filter only those rows that belong to cluster A. Since cluster A contains the players that have the best weighted average among the others. Similarly for K-Means clustering we analyse the cluster data and select two clusters that have the best players.
Application Logic	• OData service is defined to expose the data defined by the entity sets. We expose the calculation views “AGG”, “ABC” and “KMEANS” created in the previous layer using the OData services in the .xsodata file. We rename it and also define the primary key column and aggregation. • For example : "Package_name::KMEANSCV"as "KMEANS" key("playerID") aggregates always; Our application uses the XS engine to support the dynamic website.

Presentation Logic	• .xsapp file is used to tell the XS Engine in the SAP HANA server that the package contains a web app. • .xsaccess describes the entry point of the web app. • SAPUI5 web application uses HTML for defining the structure and layout of a Web document by using a variety of tags and attributes. It loads all necessary SAPUI5 libraries and starts the bootstrapping process. The content is designed and defined by using XML and JavaScript respectively. We use XML to define the structure of our inner tabs. JavaScript is used to define how the application responds to the changes and clicks done by the user. The programming logic is defined here. • We have multiple tabs which define the various Predictive and Descriptive analysis created on the basis of the Baseball data. The results that define the best players are presented in the form of graphs and tables. Users can easily navigate through the app to find the list of players. The graphs create a very user friendly interface to analyse the trend and strengths of the players.

3. Technical Interface

This section goes through the technical interface part of the application. The interface being the heart of the application has been carefully modelled to get accurate results.The modelling procedure is briefly explained in the below sections.

Procedure :

The baseball dataset after the transformation phase goes into the processing stages where a model is built around the data, to get the results on which we perform the analysis. The dataset which has been cleaned and exported into SAP HANA contains raw data without the associations. We have taken measures to model the data thereby creating associations within the data which will help us fetch accurate results from the associated data.

The procedure contains two main Predictive analysis algorithms which we have used to formulate the results. The two mian algorithms used in the implementation are ABC Clustering and K means.

Associations:

To perform the Predictive analysis algorithms we just could not simply run the implementation with the dataset, rather first we needed to carefully design the dataset or model the dataset in a way we can associate the data so as when the algorithms perform the operations on the data we get better and accurate results. These are few of the association rules followed,

The Baseball dataset was divided into three categories Batting,Pitching and Fielding and we have run our algorithms on these three fields.

Batting:

The main idea behind a batter is to get hits. He hits the ball in the opposite direction of the pitcher which helps him to get hits and in turn he scores runs. Runs being dependent on many factors, to find out the best batter who can score more runs in any circumstances, we analysed the dataset to find associations between different columns of the batter which tells us clearly that this batter can score more runs because he has either scored on these circumstances or he is capable to do so.

As mentioned the data is filtered out to contain players who have ever played a game from the past 25 years, we have created the associations by these rules,

The first rule we have used to calculate is given by,

ABHR = At Bats / Home (At Bats per Home Run)[A hit for one base is called a single]

TB = 1B + 2 * 2B + 3 * 3B + 4 * HR (1)

Here from the above equation (1) we can see that have summed up all the associated columns which are related. This was one of the association rule we have created to minimise the data complexity and also this tells us that a good ABHR rate might be one of the good factor in judging a batter.

The second rule we have used is the slugging percentage and we have calculated it by the following, Slugging percentage tells us the power of the hitter, also called Batter. Slugging percentage is calculated as Total Bases divided by At Bats.

SP: “TB"/"AB_SUM" (2)

The third rule which we used is called On Base Percentage. Which helps us identify how many times the batter can hit the ball perfectly from his base of the bat.On base Percentage is calculated as

(Hits + Base on Ball + Hit By Pitch) divided by (At Bats + Base on Balls + Hit by Pitch + Sacrifice Flies)

OBP: (float(“H_SUM")+float("HBP_SUM")+float("BB_SUM")) divided by (float(“AB_SUM") +float("HBP_SUM")+float("BB_SUM")+float("SF_SUM")) (3)

These Equations (1), (2) and (3) altogether tells us that if the batter has a good Slugging percentage, OnBasePercentage and also ABHR he can be termed as a good.

Now that we have associated the batting field of our baseball dataset which now after association contains very few columns and which can be easily sent into the predictive analysis to get our analysis results, we moved on to the second field which is Pitching.

Pitching:

An act of throwing baseball to the home plate to start the game. A pitcher would be responsible in controlling the runs scored by the batter. A good pitcher can help the team win the game. He is also an integral part of the squad.

From the Baseball dataset we have used a different association rule to model the data for the pitcher. We have firstly used this to get how effective the pitcher is.

Pitcher: Walks (BB) + Hits Divided by: Innings Pitched (4)

We then calculate the walks for 9 innings(Traditional Length of a game) by using this formula,
BB/9 = 9 x (Walks / (Innings Pitched + (Outs (Partial Innings) Pitched / 3)) (5)

Secondly, we have taken another counter measure to verify how good the pitcher is by calculating his Earn Run Average, Which will give us the mean of earn runs given by the pitcher per 9 innings.

In Baseball, Component ERA is calculated as:

Component ERA = ((((Hits + Walks + Hit by Pitch) x PTB)/(Batters Faced by Pitcher x (Innings Pitched + (Outs (Partial Innings) Pitched / 3))) x 9) (6)

Formulating all of these association rules we get a minimised pitcher data with all the associations modelled perfectly, on which we have performed our predictive analysis on. These associations have been calculated using Calculation view in SAP HANA.

Fielding:

The game contains a fielding section which help us determine a good and fit fielder can prove a valuable asset to the team. A player who is a good batter and also a good fielder, or a picther and also a good fielder adds lot of value to the team.

From the dataset we have calculated a players fielding worthiness using the below formulas.

This will give us a players fielding capability on the ground.

Fielding Percentage (FPCT) = (put outs + assists) / (put outs + assists + errors) (7)

We have also calculated the fielders total chances, so know how accurate he is on the ground.

TC = putouts + assists + errors. (8)

TC repersents Total Chance, which will give a added result of the palyers fielding capabilities.

We also calculate the fielding percentage by using this formula below.

double(“H_SUM")+double((2*"2B_SUM"))+double((3*"3B_SUM"))+double((4*"HR_SUM")) (9)

Salary:

We have also taken players salary into consideration and have modelled the data.Salary of a player also will let us know how valuable the player is in the market which in turn help us to make affordable choices.

These association rules help model the data and perform our analysis. These rules have been formulated by the help of calculation views in SAP HANA. The BaseBall dataset loaded in the SAP HANA has been thoroughly analysed and associated by the above mentioned rules. These rules formulate the backbone of our technical interface which in turn define our analysis.

4. Data Management and Data Model

Here in this section we tend to explain how we have managed the large dataset and tools we have used to model our data. We have four different csv files for Batting, Pitching, Fielding and Salary values of the different players. The data is imported from the .csv file to the editor. The data is then added to the catalog using .hdbtable and .hdbti files. The data remains in the repository until it is activated. Once activated, the data is transferred to the catalog.

- .hdbtable file defines the structure and schema of the table to be created.

For example:

```
table.schemaName ="MONEYBALL_PROJECT";
table.tableType = COLUMNSTORE;
table.columns =[
{  name="playerID"; sqlType=NVARCHAR;length=30;},
 {name= "G_SUM" ; sqlType=INTEGER;},
 {name="GS_SUM" ; sqlType=DOUBLE;},
{name=" InnOuts_SUM" ; sqlType=DOUBLE;}
];
```

- .hdbti file defines the table name, schema name in the catalog where the table has to be exported. It also connects the hdbtable to the csv file.

```
import = [
        {
            table = "gbi-student-015.MoneyBall::Pitching" ;
            schema = "MONEYBALL_PROJECT";
            file = "gbi-student-015.MoneyBall:Pitching.csv";
            header = false;
        }
      ];
```

Calculation view:

Once the data is added to the catalog, we create a calculation view of type CUBE with star join. This calculation view joins the different tables with respect to the primary key, “playerID” to get a master table. We use various formulas to get the different measures that evaluate a player’s capability.

- Slugging percentage was calculated as (Total Bases / At Bats).
- On base Percentage was calculated
- Equivalent Average is calculated
- Fielding Percentage
- We then add Filter conditions to get rid of the divide by zero errors and other irregularities.

Finally a weighted average was calculated by taking all the various measures into consideration.

When the calculation view is saved we get a new table with cleaned data.

SAP Expert Analytics Tool:

The data is then imported into the SAP Expert Analytics tool to perform the predictive analysis.

ABC clustering: The data is added to the ABC clustering algorithm and given appropriate percentages for each. The clustering was done on the basis of the “Weighted average” column. The outcome shows that the cluster A contains the players with the highest weighted average.

K-Means clustering: The K-means algorithm makes use of “Weighted average” column for clustering. We have considered 5 clusters in this scenario. The outcome shows that clusters 1 and 4 contain the best players.

The result of this analysis is written back into SAP HANA using the data writer.

SAP UI5:

The processed data is then presented to the users using the SAP UI5 app. We bind the data into the app using the OData services.

For example:

"Package_name::ABCSAP"as "ABC"

key("playerID")

aggregates always;

SAPUI5 web application uses index.html file to define the structure and loads all necessary SAPUI5 libraries and starts the bootstrapping process. The content is designed and defined by using XML and JavaScript respectively.

The overview tab shows the current salaries and weighted average score of the players. The ABC and K-Means cluster predictive analysis is depicted using graphs and tables. We have then displayed descriptive Batting, Pitching and Salary analysis in three other graph structures. These graphs create a very user friendly interface to analyse the trend and strengths of the players.

Data Security requirements

Data security has been one of the prime concerns and enough measures have been taken in place to ensure the security of the application. Measures such as only the team developers have the full access to the application and the rest of the users have certain access restrictions have been deployed. The security measures are further addressed in the document in the below section.

Data Archiving

Removal of unwanted data from the dataset is called archiving. There are various steps taken in place to archive the data which is not used. The old data which is of no use to the prediction analysis have been carefully identified and have been removed. This process of removal of unwanted data is called archiving. In our dataset players whose data are more than 25 years old are termed as old and have been achieved.

5. Functional Properties of the System

5.1 Staging Layer Load

5.1.1 Overview

The changes made to the SAP HANA database (Which now contains the baseball data)such as insert, delete, update etc, needs to be backed up into the backup node. These changes are logged into the staging store by using the information from the log parser.

5.1.2 Variables and Parameters

Batting Variables:

G_SUM,AB_SUM,R_SUM,H_SUM,2B_SUM
3B_SUM,HR_SUM,RBI_SUM,SB_SUM,CS_SUM,BB_S,M,SO_SUM,IBB_SUM,HBP_SUM,SH_SUM, SF_SUM, GIDP_SUM

Fielding Variables:

G_SUM, GS_SUM, InnOuts_SUM, PO_SUM, A_SUM, E_SUM, DP_SUM

Pitching variables:

SO_SUM,BAOpp_SUM,ERA_SUM,IBB_SUM,WP_SUM,HBP_SUM,BK_SUM,BFP_SUM,GF_SUM,R_SUM,SH_,UM,SF_SUM,GIDP_SUM,W_SUM,L_SUM,G_SUM,GS_SUM,CG_SUM,SHO_SUM,SV_SUM,IPouts_SUM,H_SU,,ER_S, UM,HR_SUM,BB_SUM.

Parameters:

On-Base Percentage, Sluggish Percentage, Fielding_Percentage, WalksPer9Inning, WalkToStrikeOutRatio, EQA, BattingAvg, BABIP.

The expansions for these abbreviations can be found in the readme documentation.

5.1.3 Database Entities

Batting table, Fielding table, Pitching table , Salary table

5.1.4 Pre-Execution Processes

We make sure that there are no errors in the process and write statements needed for update, delete and join.

5.1.5 Post-Execution Processes

Post execution an sql file and table is created that contains the data necessary for our analysis. In our case, only the clusters that contain players with the highest weights average will be populated.0

5.1.6 Transformation Details

The tables are aggregated to create a master table that contains all the data. The rows with empty columns or corrupted data is deleted and cleaned. The filter conditions are applied to get the important data required for the processing.

For Predictive analysis using ABC clustering, we have filtered only those rows that beong to cluster A. Since cluster A contain the players that have the best weighted average among the others. Similarly for K-Means we analyse the cluster data and select two clusters that have the best players.

5.2 Corporate Store Load

5.2.1 Process Description

Users who need access to the project (Users who wish to view our analysis or a developer who wants to improve the application) will be given unique user credentials which will be authenticated on the cloud platform.

5.2.2 Variables and Parameters

User credentials / parameters such as User ID and password is required to log in to the application

5.2.3 Database Entities

UserDetails, Batting table, Fielding table, Pitching table , Salary table

5.2.4 Pre-Execution Processes

Each user who needs access to the application has to be provided with unique user ID’s. The users are then allowed to set their passwords.

5.2.5 Post-Execution Processes

The user logs out from the application once the process is complete.

5.2.6 Process Description

The users are first given the credentials needed to login to the application. Once the credentials are set, if the authentication is successful, the user can log in into the application. The session can be closed by logging out of the application.

6. Non Functional Requirements

6.1 Performance

With ever increasing markets the performance plays a vital role be it in any field. Here performance can be defined in many ways. Once such way being the performance of the analysis or the results with our dataset. How well the application was able to process the data which we modelled and was able to generate results which in turn helped the client. How was the performance of our data model? Was it good enough to able to generate good results? These define performance of our applications. Performance of our data model has shown good results for the dataset which was given.

6.2 Availability

The data which we obtained after the implementation process, how much of it was used? how much of it was reliable and useful information to the client or the end user where in he could base his decisions off? that defines what availability means here. The data model created processed with the predictive analysis how much data was generated so that the generated data was useful. As the dataset grows the availability of the data change. Here in our implementation the data generated after the implementation was highly reliable and useful data.The quality of data generated and was of greater use was more.

6.3 Maintainability, Adaptability and Portability

As and when the data grows, we need certain mechanisms to be in place which can adapt to the change in circumstances. Our model of the data has to adapt, maintain and have certain features which can handle certain sudden change in circumstance. This is where these non functional requirements come into the picture. Our data model must have the capacity to adapt to the change of the ever-growing data, must try to maintain its performance and should not cause unexpected errors when the platform on which it resides changes.

7. Security

7.1 Communication channels

Security is one of the biggest processes any organisation creating any application should consider. There are many types of security concerns one should take into account. By default, the communication channels are not secure. To secure them, we use certain protocols and channels such as Secure Sockets Layer (SSL) protocol. SSL is a cryptographic protocol that provides security and data integrity for communications over TCP/IP networks. The way SSL works is by the property of handshaking among client and server and their by initiating the further processes.

7.2 Application-specific security

The application access is restricted to only valid users having a “Valid User Identification” and “Password”. They can login to the application with their credentials and can perform any valid operations on the data (such as read, write and delete), more often the developers who are working on the project will be using or have been granted (all) read, write and the delete operations to cope up with the needs of the requirements. Any members other than the project members who are involved indirectly to the project would have less access and cannot perform all the operations and would be advised to use the application for analysis or evaluation purposes only.

Access Rights and Roles	Read	Write	Delete
Developer	X	X	X
Business Owner	X		
Others	X		

8. References

[1]. SAP TDD Template,Shared by the professor Stephan.Willi hart.

[2] Lehman’s Database - http://seanlahman.com/baseball-archive/statistics (Version: 2016)