**Tech Challenge – Devoteam Red Sox**

This document will serve as a storage for all design decisions, assumptions I made with their rational, or other notes that should be kept for the presentation.

**Basic information**

Our clients are the **Red Sox**, and they want help with their **game analysis**. There is a massive amount of data, but they don’t know where to start to learn something.

The goal is to start from the subset of data they shared with you to show them:

* How to get data into BigQuery
* How to structure data in BigQuery
* How to visualize the data with Looker Studio

Make sure to make **components scalable for future growth of the platform** (!)

**Must haves**

* Create at least 10 tables in BigQuery. You may use different modelling layers or modelling techniques as you see fit
* Make sure the setup can incorporate new data in a batch fashion.
* Make sure the setup is scalable & reusable to more data and use cases.
* Make at least 1 dashboard visualizing insights into the data in Looker Studio.
* Allow end users to do some exploration themselves on the shown data in Looker Studio.

**Nice to haves**

* Make it possible to restore a backup of the platform to recover from problems.
* Make dashboards real-time (less than 2 min delay on newly arrived data).
* Make sure you get an alert if something fails.
* Make a setup that allows for easy testing before going to production.
* …

**Plan**

1. ~~Explore the data (1 dag)~~
2. Background research
   1. ~~Learn basic baseball operation~~
   2. Learn baseball specific metrics
3. ~~Plan tables~~
   1. ~~Design data model~~
4. ~~Research techniques for the requirements~~
   1. ~~Scalability techniques~~
   2. Batch technique
   3. ~~Reusability techniques (draw pipeline overview?)~~
5. ~~Create tables (/views/materialized views?)~~
6. ~~Plan visuals~~
   1. ~~User interaction (filtering, sorting, dropdowns, drill-downs?)~~
7. ~~Create visuals~~
8. ~~Use and check functionality (test filters, reloads, data connection?)~~
9. ~~Document~~
10. ~~Share project with the required people~~
11. Prepare interview

**Explore the data**

The dataset is the **bigquery\_public\_data.baseball** dataset.

The dataset consists of 3 tables: **schedules**, **games\_wide**, **game\_post\_wide**:

* **Games\_wide** consists of 145 columns, 761 618 rows. Table schema can be seen in bigquery’s schema tab.

The number of rows is very big because the table records all events happening during games for the entire season

* **Games\_post\_wide** is the same as games\_wide but for the post season. It has the exact same schema. It consists of 8676 rows.
* **schedules** contains information on the planning of regular season games. It consists of 16 columns and 2431 rows representing one game each.

**Background research**

Before deciding on what tables to setup etc. I want to learn something about the most important statistics in baseball.

* Batting average (AVG) =
  + Hit = batter strikes the ball into fair territory and reaches base without doing so via an error or a fielder’s choice. Can be singles, doubles, triples & home runs.
  + Average over seasons is .250
  + Veel speciale gevallen tellen niet als at bat
* On base percentage (OBP) =
  + Times on base include hits, walks and hit by pitches, but do not include errors, times reached on a fielder’s choice or a dropped third strike
  + Sacrifice blunts removed entirely
* Slugging percentage (SLG) = average base reached by a batter per at bat
  + =
* On base percentage + Slugging percentage (OPS) = OBP + SLG
  + Represents how well a hitter can reach base + how well he can hit for average and for power
* On base percentage + Slugging percentage (OPS+) = 100\*(OBP / lgOBP + SLG/lgSLG -1)
  + Normalizes OBP across the league to account for external factors
  + Always has 100 as average
  + Takes into consideration a batter’s home ball park (because of size differences, altitudes, etc.)
  + Way to compare players’ OPS
* Weighted on base average (wOBA) =
  + There are different ways to get on base (single, double, triple, homerun), normal OBP doesn’t differentiate that
  + If bases are loaded (have players on them) also more impactful: taken into account
* Weighted runs above average (wRAA) = \*(AB+BB+HBP+SF+SH)
  + How much better/worse a player is at scoring runs
* Runs Created (RC) = = OBP \* SLG \*AB = OBP \* TB (total bases)
  + How many runs a batter created
* Weighted Runs Created Plus (wRC+) = (((wRAA per PA) + (league runs per PA – ballpark factor x league runs per PA) / league wRC per plate appearance, not including pitchers)) x 100
  + runs created but normalized
* Wins above replacement (WAR)

Stat Betekenis Hoe berekend

G Games gespeeld als pitcher Aantal unieke gameIds waarin pitcherId = spelerId

IP Innings pitched SUM(outs) / 3 (alle outs waarbij deze speler de pitcher was)

H Hits toegestaan Aantal events met is\_hit = 1 én pitcherId = spelerId

HR Home runs toegestaan Aantal hits waar hitType = 'home\_run'

BB Walks (vier wijd) toegestaan Aantal events waar balls = 4 én is\_ab\_over = 1

SO Strikeouts (3 strikes) Aantal events waar strikes = 3 én is\_ab\_over = 1

ERA Earned Run Average (Earned Runs / IP) \* 9 (je moet earned runs bepalen via outcomes of score progression)

WHIP Walks + Hits per Inning Pitched (BB + H) / IP

AVG Against Slaggemiddelde tegen de pitcher Hits / At Bats (waar is\_ab = 1 en pitcherId = spelerId)

G is sum gameId per pitcherId

H (Hits against) sum (is\_hit per pitcherid)

HR (Homeruns against) sum(outcomeId = 8 per pitcherId)

BB (Walks) sum(outcomeId = 24 of 26)per pitcherId

SO (StrikeOuts) sum(outcomeId = 37 of 28)per pitcherId

AVG Against Hits / At bats (At bats = is\_ab = 1 en outcomeId niet in 24 26en 7)

WHIP = (Walks + Hits) / IP

ERA

IP = Outs/3

AVERAGE SPEED

Pitch type usage

Pitch zone Usage