Soccer Data Warehouse Presentation Deck

By: Gary Buckley, Scott Kalich, Brennan Tolman, Zach Zhang
Group 9

Background

Vision & Objective

Deliver effective team game strategies ideas to the organization based on data. To help organization create and implement efficient game tactics and achieve greater business benefits.

Product & Service

SQL, Talend, DBMS, BI…

Current Scenario of the Business

Creating a data warehouse for better analysis and discover the relations of the data. To meet the objectives, the data warehouse serves to explore the effect of the locations on the goal rate.

Design team tactics based on scientifically proven information. It can improve the pertinence of training and the effectiveness in actual competition.

Challenges

- Data sets have outliers and unintuitive structure
- Large amount of data stored in 2 sets
- Unable to easily query questions based on items such as location

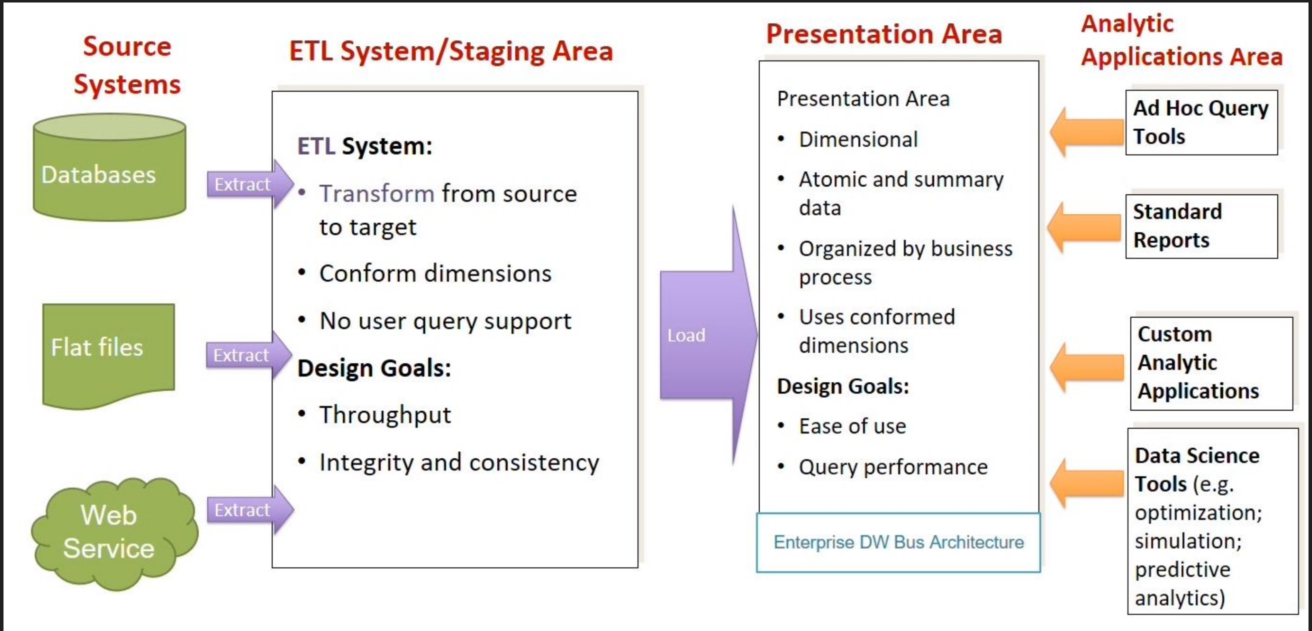
Solution

- Remove outliers prior to the data warehouse and use Talend to split tables
- Split data sets in Talend to fit a dimensional model
- Use ETL processes to create and label zones to easily slice data

Why a Data Warehouse?

- Connects multiple data sources together for analysis
- Creates consistent quality and information for end-users
- Allows a security hierarchy to block unwanted writes/updates to data
- Standardized naming conventions
- Increases ease of queries
- Decreases latency between query and result

Create a dataflow that supports end-users



Source Systems

Databases

Flat files

Web Service

Extract

Extract

Extract

ETL System/Staging Area

ETL System:

- Transform from source to target
- Conform dimensions
- No user query support

Design Goals:

- Throughput
- Integrity and consistency

Load

Presentation Area

Presentation Area

- Dimensional
- Atomic and summary data
- Organized by business process
- Uses conformed dimensions

Design Goals:

- Ease of use
- Query performance

Enterprise DW Bus Architecture

Analytic Applications Area

Ad Hoc Query Tools

Standard Reports

Custom Analytic Applications

Data Science Tools (e.g. optimization; simulation; predictive analytics)

Business Requirements

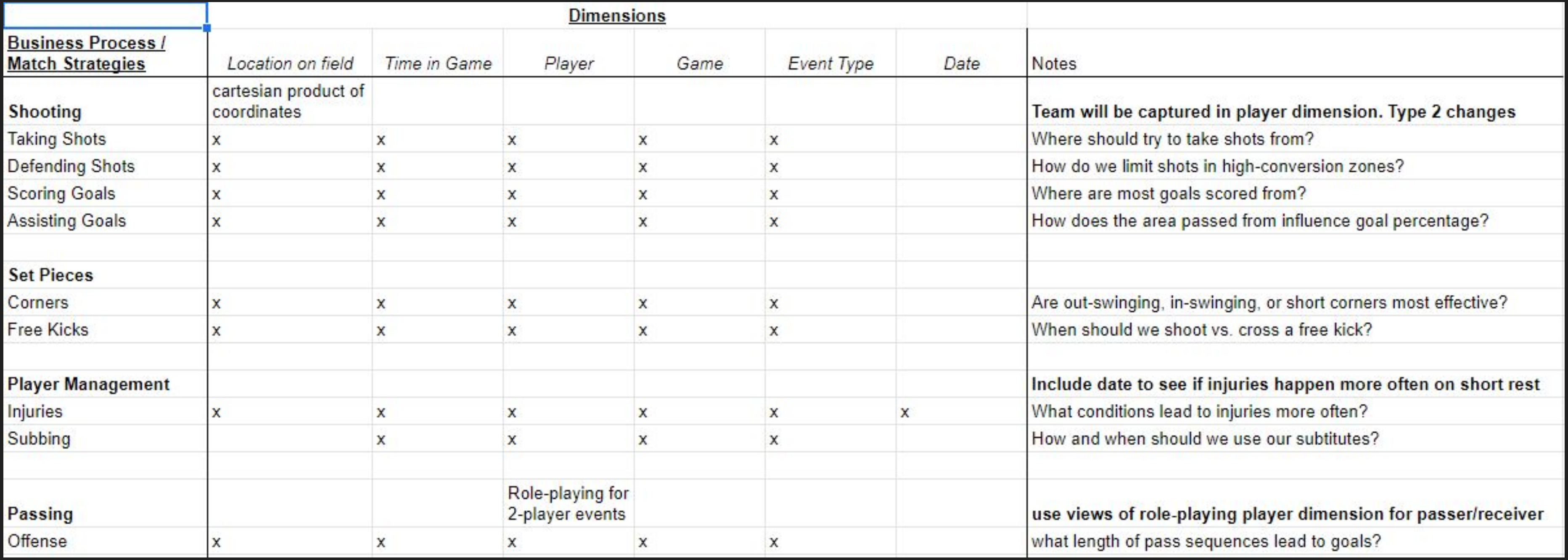
Current Requirements

- Where are most shots generated from?
- Where are most goals generated from?
- What time of the game yields the most shots and goals?
- Which players lead the league in shots and goals?

Potential Future Requirements

- Where do key passes (passed that lead to shots) come from?
- Where do assists come from?

Bus Matrix



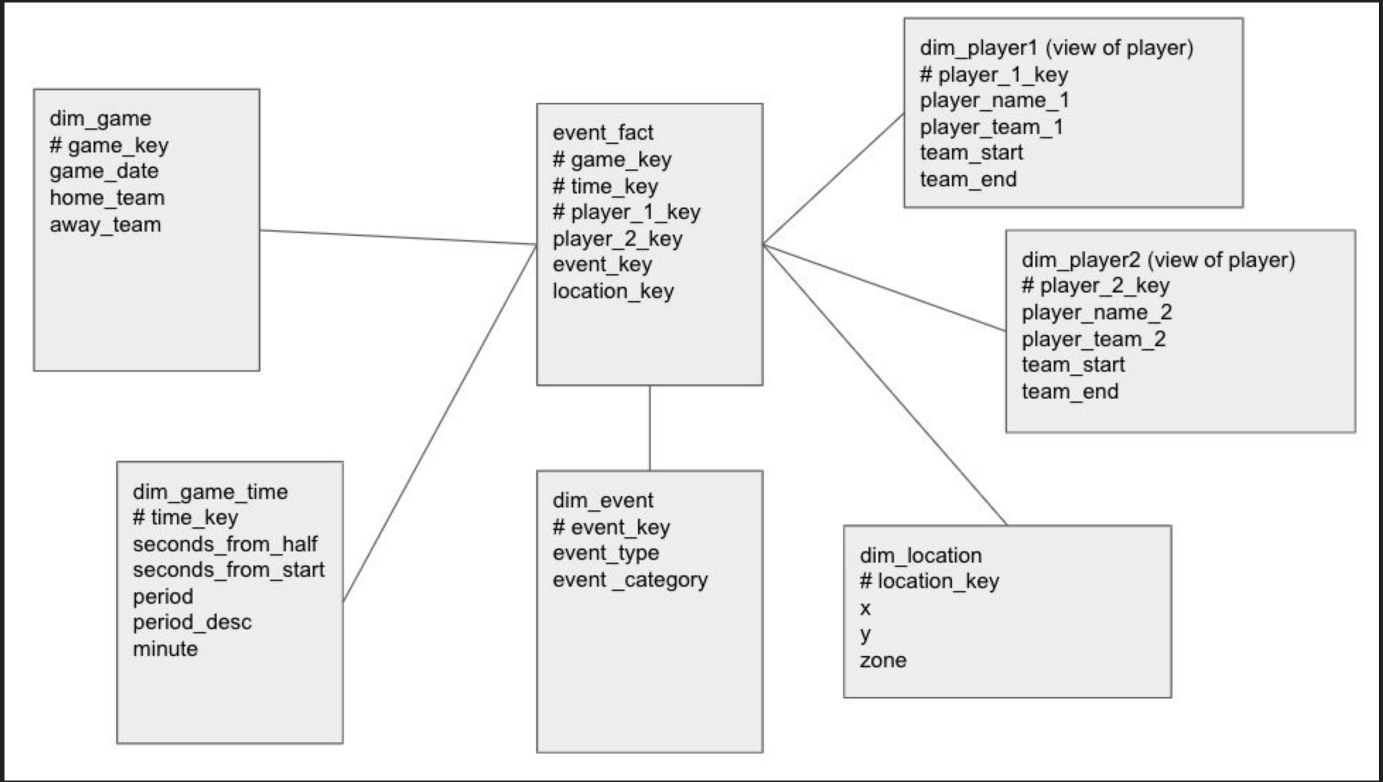
	Dimensions						
Business Process / Match Strategies	*Location on field*	*Time in Game*	*Player*	*Game*	*Event Type*	*Date*	Notes
Shooting	cartesian product of coordinates						**Team will be captured in player dimension. Type 2 changes**
Taking Shots	x	x	x	x	x		Where should try to take shots from?
Defending Shots	x	x	x	x	x		How do we limit shots in high-conversion zones?
Scoring Goals	x	x	x	x	x		Where are most goals scored from?
Assisting Goals	x	x	x	x	x		How does the area passed from influence goal percentage?
Set Pieces							
Corners	x	x	x	x	x		Are out-swinging, in-swinging, or short corners most effective?
Free Kicks	x	x	x	x	x		When should we shoot vs. cross a free kick?
Player Management							**Include date to see if injuries happen more often on short rest**
Injuries	x	x	x	x	x	x	What conditions lead to injuries more often?
Subbing		x	x	x	x		How and when should we use our subtitutes?
Passing			Role-playing for 2-player events				**use views of role-playing player dimension for passer/receiver**
Offense	x	x	x	x	x		what length of pass sequences lead to goals?

Dimensional Model Approach

General Strategy

- Dimension tables are created based on attributes
- Attributes uniquely identify instances on dimensions
- Dimensions link together to a fact table
- Fact table holds frequently queried information

Dimensional Model



dim_game
# game_key
game_date
home_team
away_team

event_fact
# game_key
# time_key
# player_1_key
player_2_key
event_key
location_key

dim_player1 (view of player)
# player_1_key
player_name_1
player_team_1
team_start
team_end

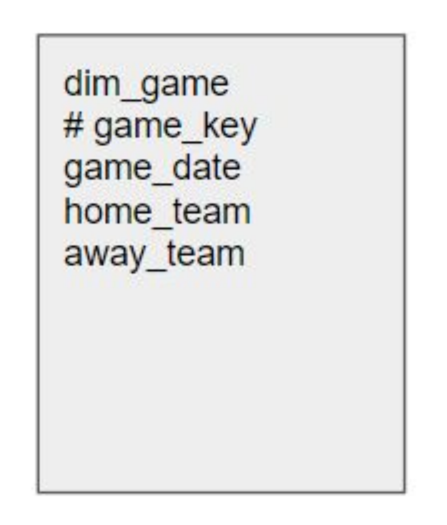
dim_player2 (view of player)
# player_2_key
player_name_2
player_team_2
team_start
team_end

dim_game_time
# time_key
seconds_from_half
seconds_from_start
period
period_desc
minute

dim_event
# event_key
event_type
event _category

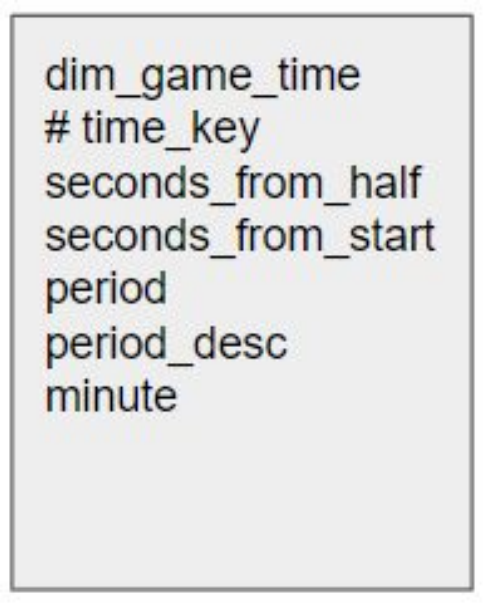
dim_location
# location_key
x
y
zone

Game Dimension



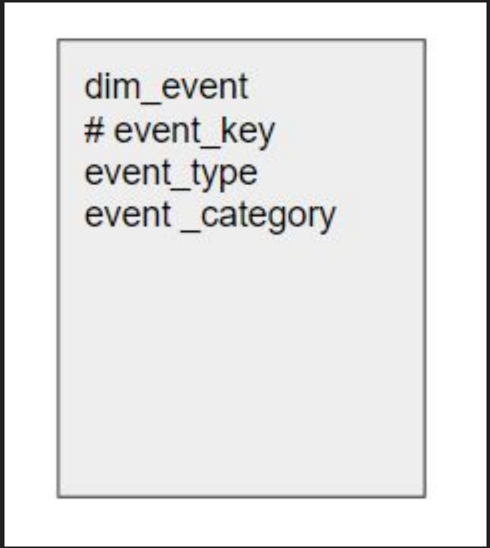
dim_game
# game_key
game_date
home_team
away_team

Game Time Dimension



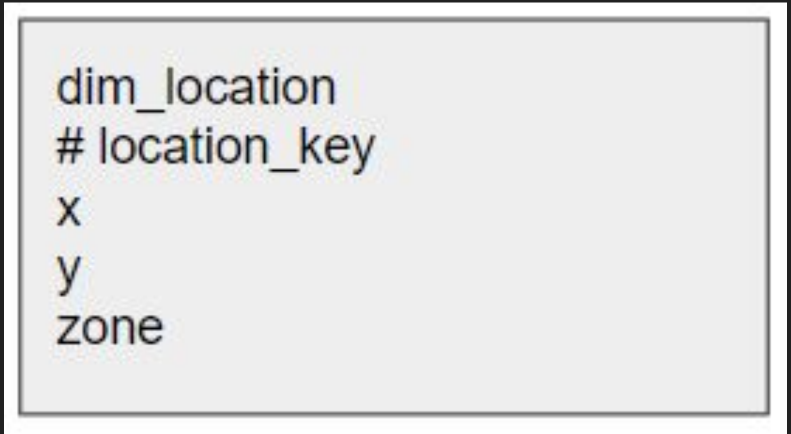
dim_game_time
# time_key
seconds_from_half
seconds_from_start
period
period_desc
minute

Event Dimension (modified)



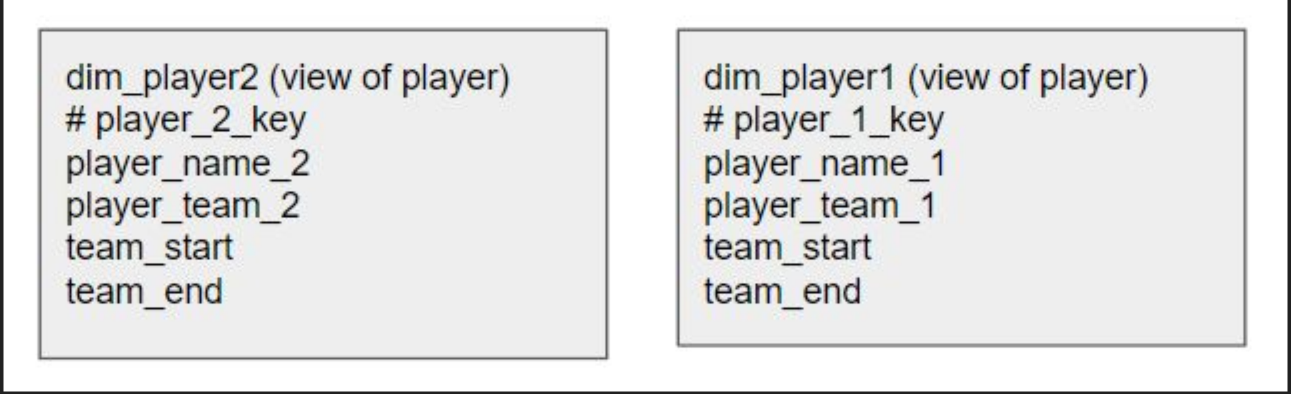
dim_event
# event_key
event_type
event _category

Location Dimension (modified)



dim_location
# location_key
x
y
zone

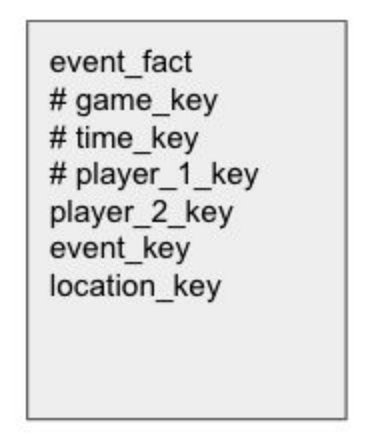
Player Dimension (role-playing)



dim_player2 (view of player)
# player_2_key
player_name_2
player_team_2
team_start
team_end

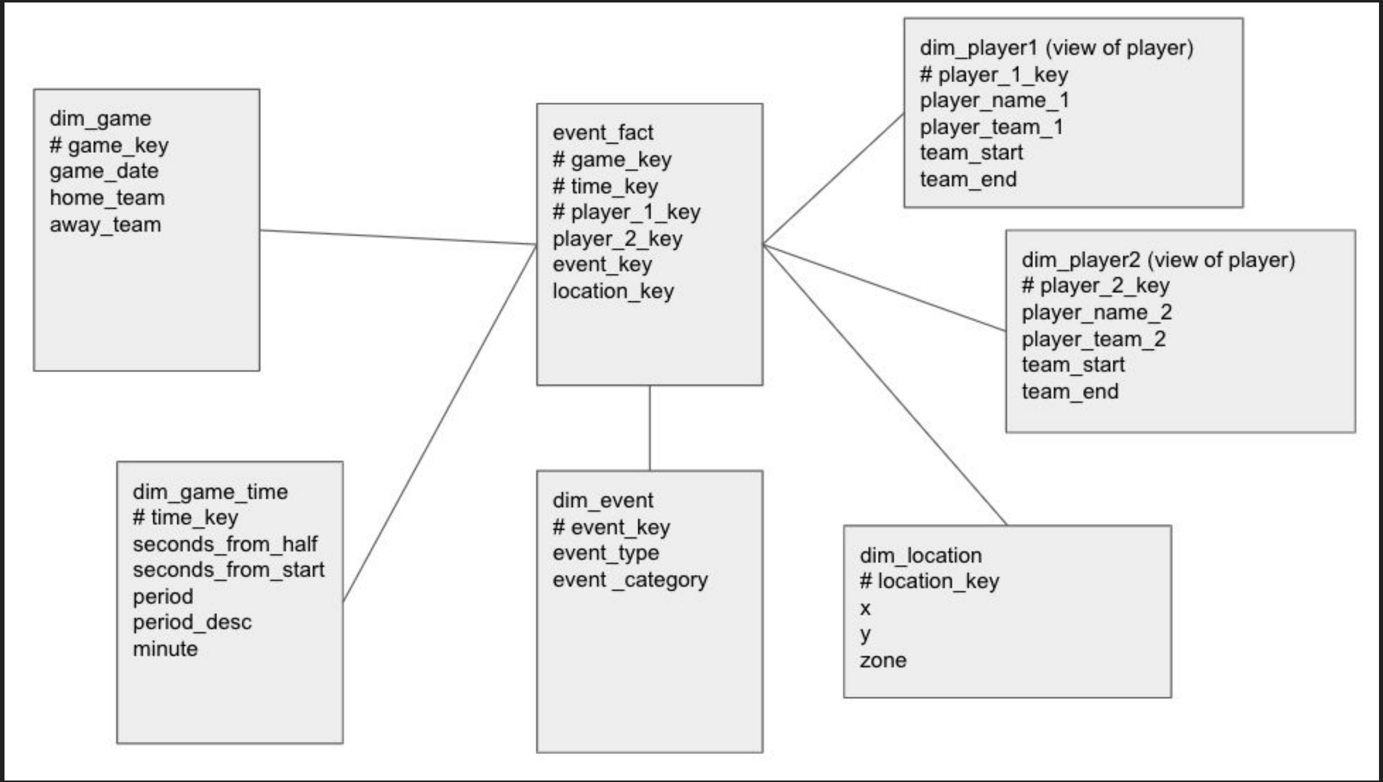
dim_player1 (view of player)
# player_1_key
player_name_1
player_team_1
team_start
team_end

Event Fact Table



```
event_fact
# game_key
# time_key
# player_1_key
player_2_key
event_key
location_key
```

Dimensional Model



dim_game
# game_key
game_date
home_team
away_team

event_fact
# game_key
# time_key
# player_1_key
player_2_key
event_key
location_key

dim_player1 (view of player)
# player_1_key
player_name_1
player_team_1
team_start
team_end

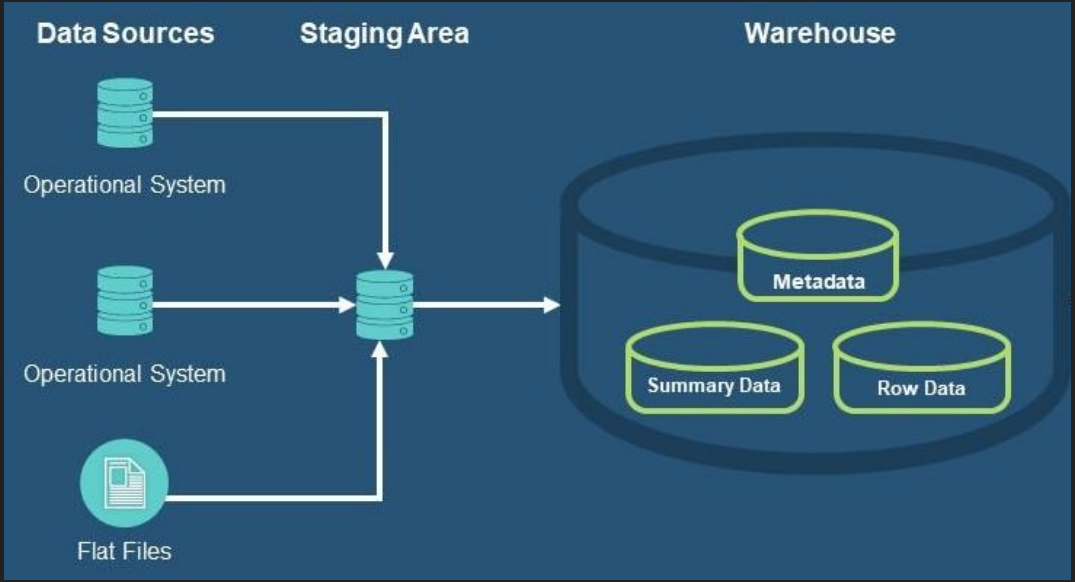
dim_player2 (view of player)
# player_2_key
player_name_2
player_team_2
team_start
team_end

dim_game_time
# time_key
seconds_from_half
seconds_from_start
period
period_desc
minute

dim_event
# event_key
event_type
event _category

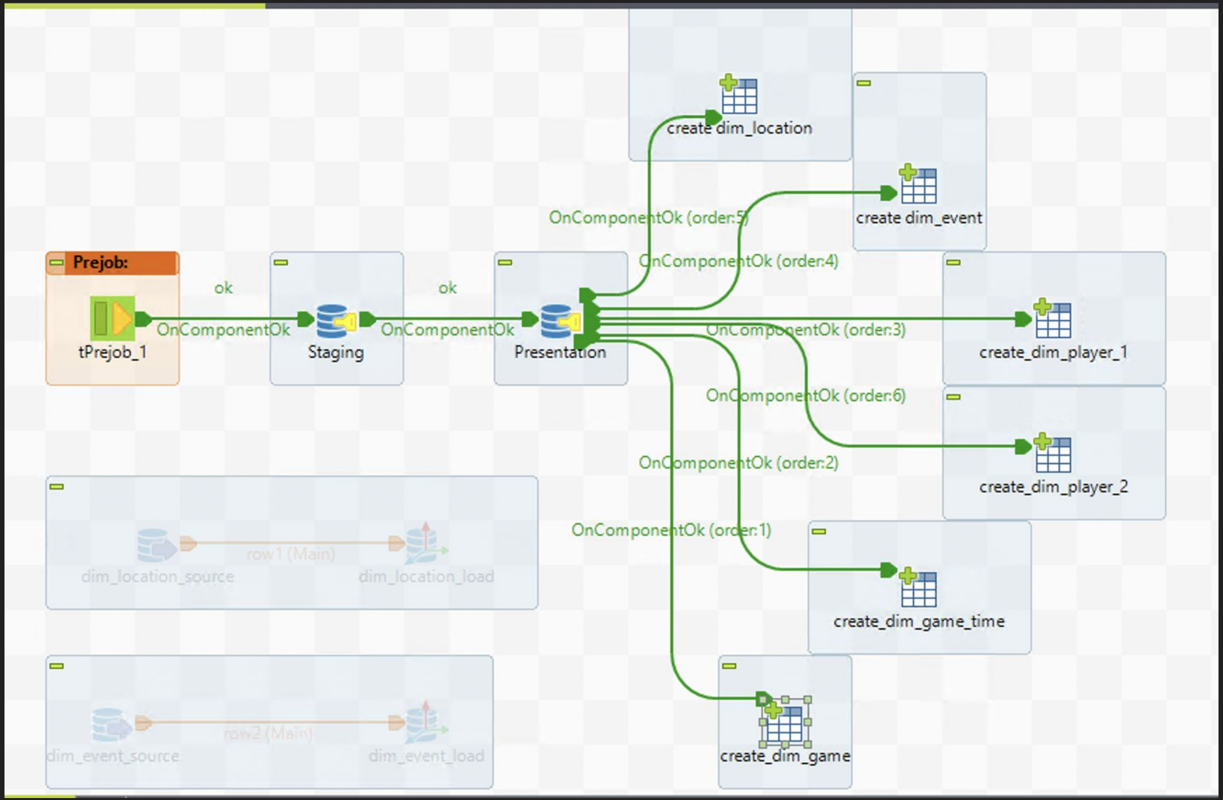
dim_location
# location_key
x
y
zone

Extract-Transform-Load

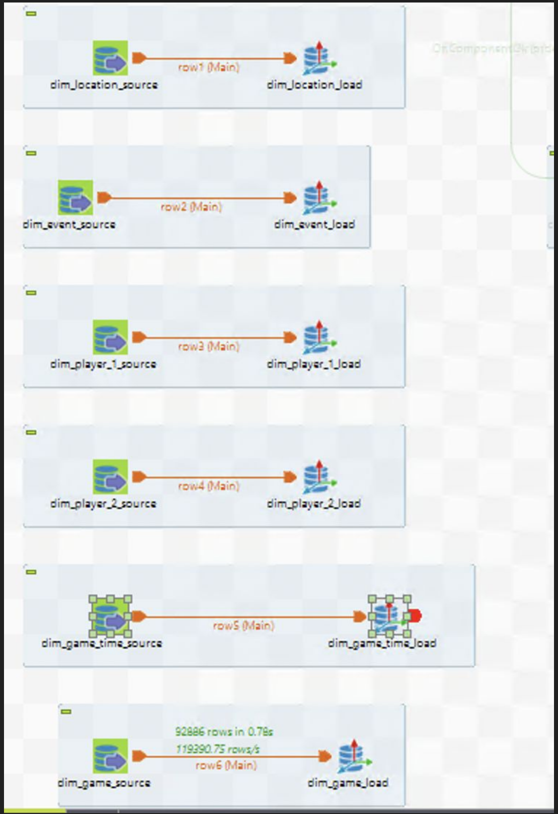


- Grab data from sources (flat file)
- Transform data by removing outliers, noise, and unused data
- Load clean data into warehouse to be used

Dimension Creation

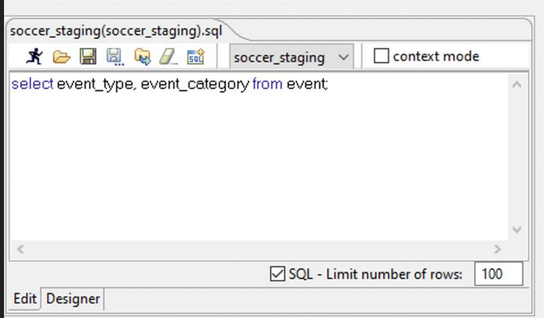


Data Loading



92886 rows in 0.78s
119390.75 rows/s

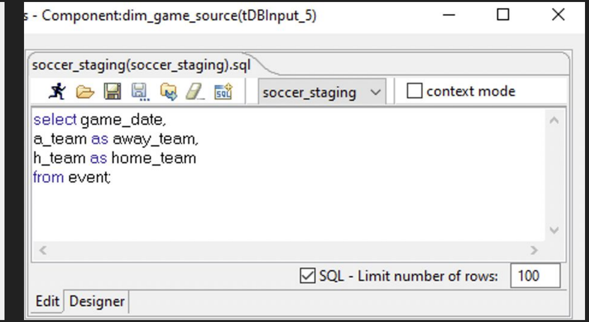
Queries



soccer_staging ☐ context mode

```sql
select event_type, event_category from event;
```

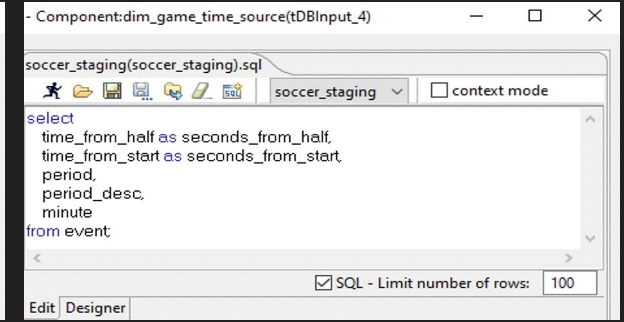
☑ SQL - Limit number of rows: 100


s - Component:dim_game_source(tDBInput_5)
soccer_staging(soccer_staging).sql
Edit
Designer

soccer_staging ☐ context mode

```sql
select game_date,
a_team as away_team,
h_team as home_team
from event;
```

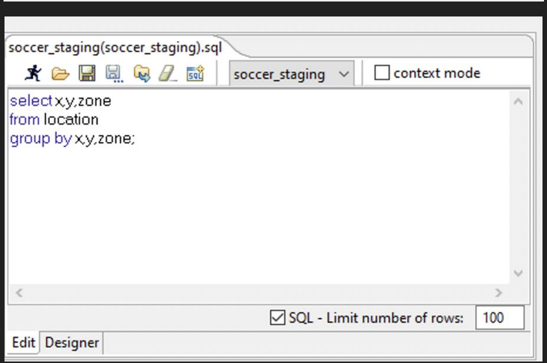
☑ SQL - Limit number of rows: 100


- Component:dim_game_time_source(tDBInput_4)
soccer_staging(soccer_staging).sql
Edit
Designer

soccer_staging ☐ context mode

```sql
select
   time_from_half as seconds_from_half,
   time_from_start as seconds_from_start,
   period,
   period_desc,
   minute
from event;
```

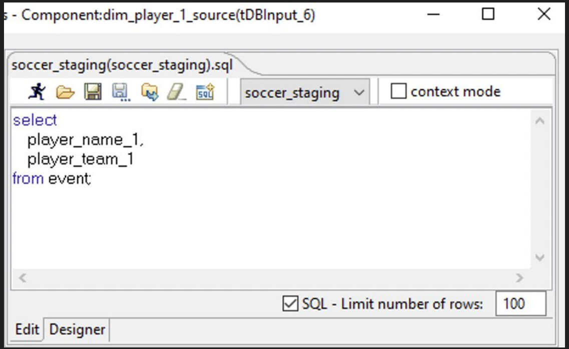
☑ SQL - Limit number of rows: 100



soccer_staging ☐ context mode

```sql
select x,y,zone
from location
group by x,y,zone;
```

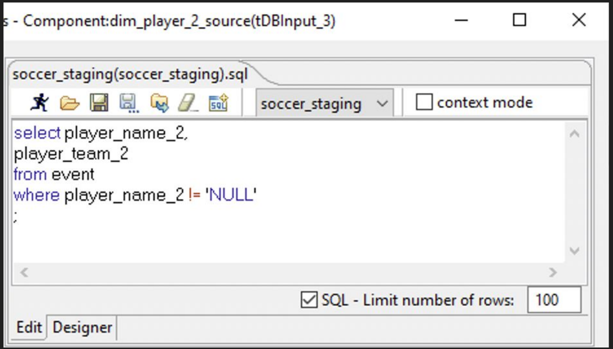
☑ SQL - Limit number of rows: 100


s - Component:dim_player_1_source(tDBInput_6)
soccer_staging(soccer_staging).sql
Edit
Designer

soccer_staging ☐ context mode

```sql
select
   player_name_1,
   player_team_1
from event;
```

☑ SQL - Limit number of rows: 100


s - Component:dim_player_2_source(tDBInput_3)
soccer_staging(soccer_staging).sql
Edit
Designer

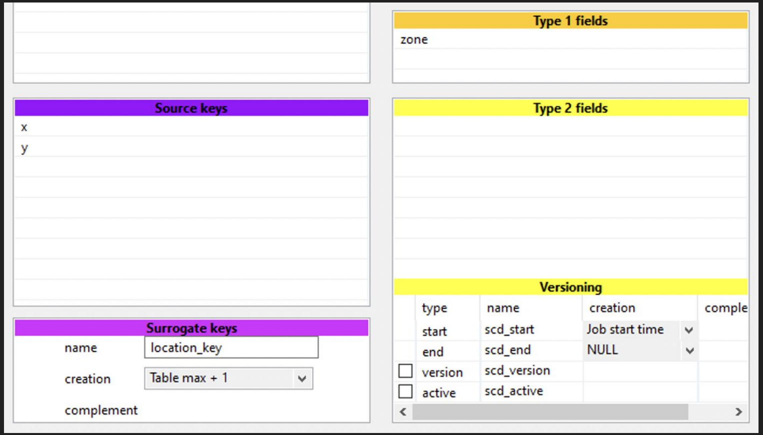
soccer_staging ☐ context mode

```sql
select player_name_2,
player_team_2
from event
where player_name_2 != 'NULL'
;
```

☑ SQL - Limit number of rows: 100

Keying

Location



Type 1 fields

zone

Source keys

x

y

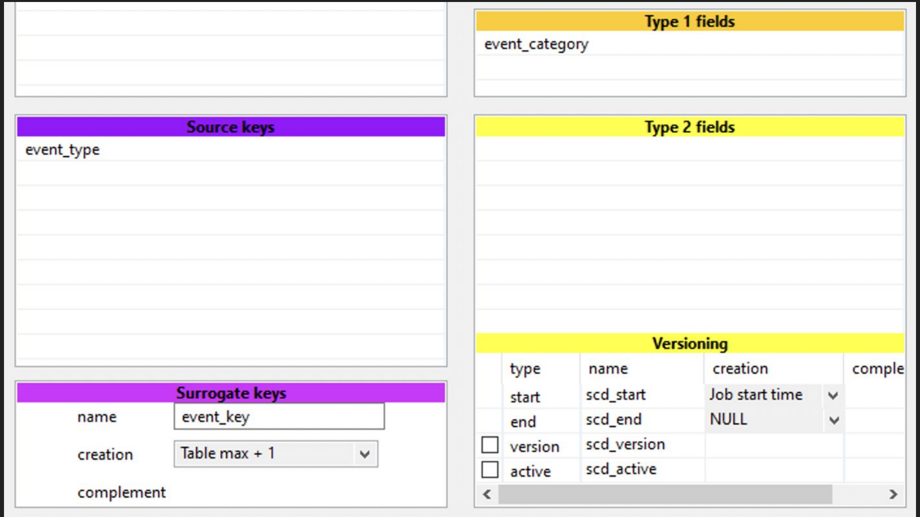
Type 2 fields

Versioning

type	name	creation	comple
start	scd_start	Job start time	
end	scd_end	NULL	
□ version	scd_version		
□ active	scd_active		

Surrogate keys

name	location_key
creation	Table max + 1
complement	



Type 1 fields

event_category

Source keys

event_type

Type 2 fields

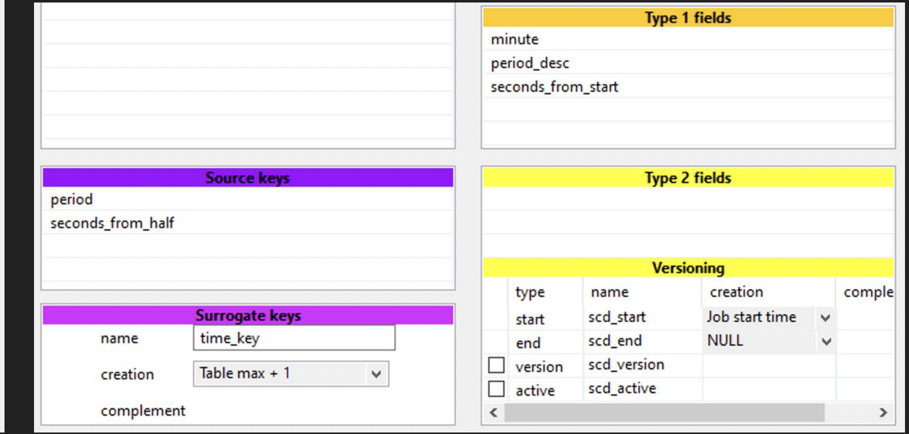
Versioning

type	name	creation	comple
start	scd_start	Job start time	
end	scd_end	NULL	
□ version	scd_version		
□ active	scd_active		

Surrogate keys

name	event_key
creation	Table max + 1
complement	

Event



Type 1 fields

minute

period_desc

seconds_from_start

Source keys

period

seconds_from_half

Type 2 fields

Versioning

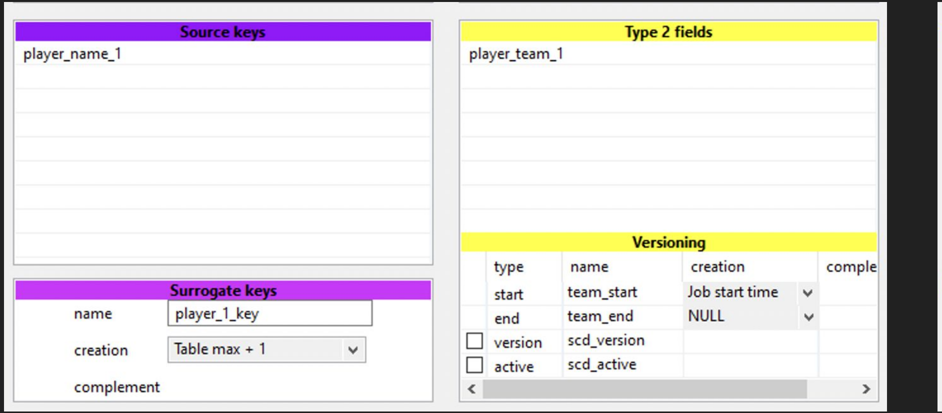
type	name	creation	comple
start	scd_start	Job start time	
end	scd_end	NULL	
□ version	scd_version		
□ active	scd_active		

Surrogate keys

name	time_key
creation	Table max + 1
complement	

Time

Keying



Source keys

player_name_1

Surrogate keys

name	player_1_key
creation	Table max + 1
complement	

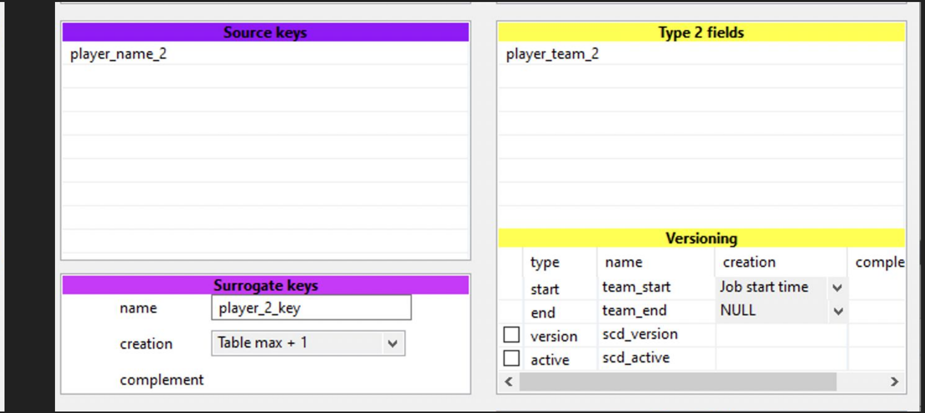
Type 2 fields

player_team_1

Versioning

	type	name	creation	comple
	start	team_start	Job start time	
	end	team_end	NULL	
□	version	scd_version		
□	active	scd_active		

Player 1



Source keys

player_name_2

Surrogate keys

name	player_2_key
creation	Table max + 1
complement	

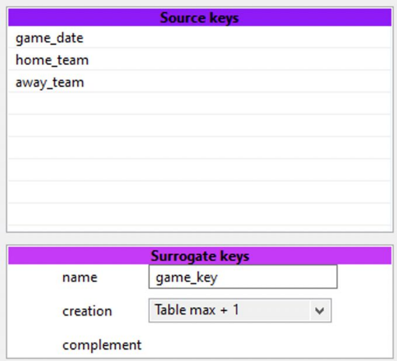
Type 2 fields

player_team_2

Versioning

	type	name	creation	comple
	start	team_start	Job start time	
	end	team_end	NULL	
□	version	scd_version		
□	active	scd_active		

Player 2



Source keys

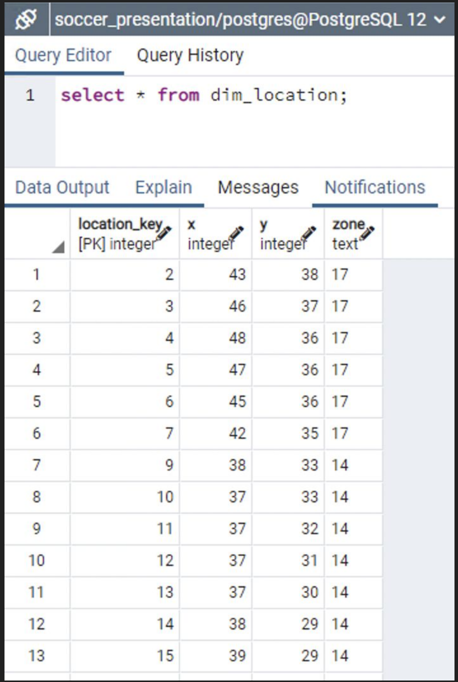
game_date
home_team
away_team

Surrogate keys

name	game_key
creation	Table max + 1
complement	

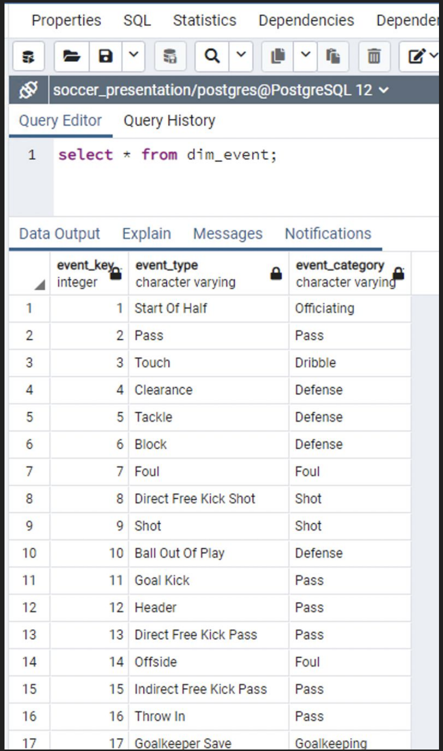
Game

Sample Data



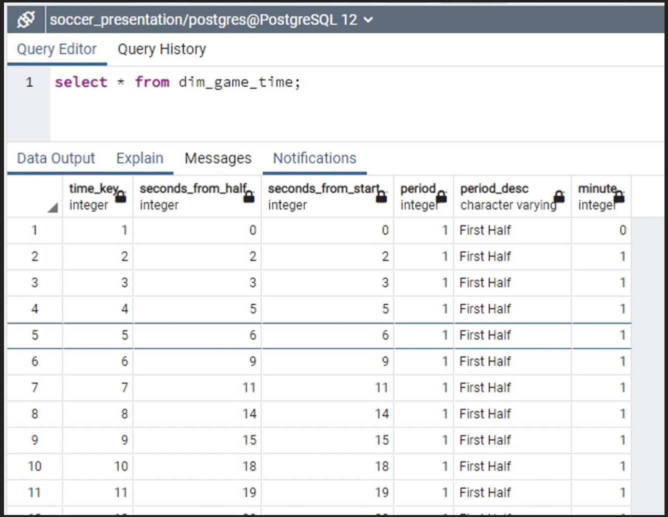
```
select * from dim_location;
```

	location_key [PK] integer	x integer	y integer	zone text
1	2	43	38	17
2	3	46	37	17
3	4	48	36	17
4	5	47	36	17
5	6	45	36	17
6	7	42	35	17
7	9	38	33	14
8	10	37	33	14
9	11	37	32	14
10	12	37	31	14
11	13	37	30	14
12	14	38	29	14
13	15	39	29	14



```
select * from dim_event;
```

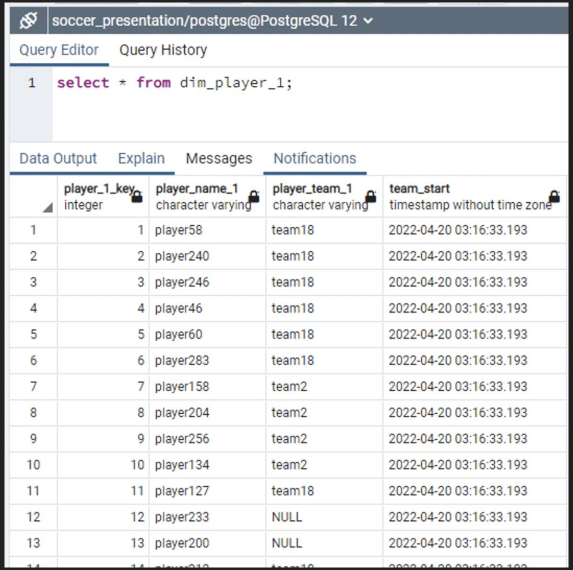
	event_key integer	event_type character varying	event_category character varying
1	1	Start Of Half	Officiating
2	2	Pass	Pass
3	3	Touch	Dribble
4	4	Clearance	Defense
5	5	Tackle	Defense
6	6	Block	Defense
7	7	Foul	Foul
8	8	Direct Free Kick Shot	Shot
9	9	Shot	Shot
10	10	Ball Out Of Play	Defense
11	11	Goal Kick	Pass
12	12	Header	Pass
13	13	Direct Free Kick Pass	Pass
14	14	Offside	Foul
15	15	Indirect Free Kick Pass	Pass
16	16	Throw In	Pass
17	17	Goalkeeper Save	Goalkeeping



```
select * from dim_game_time;
```

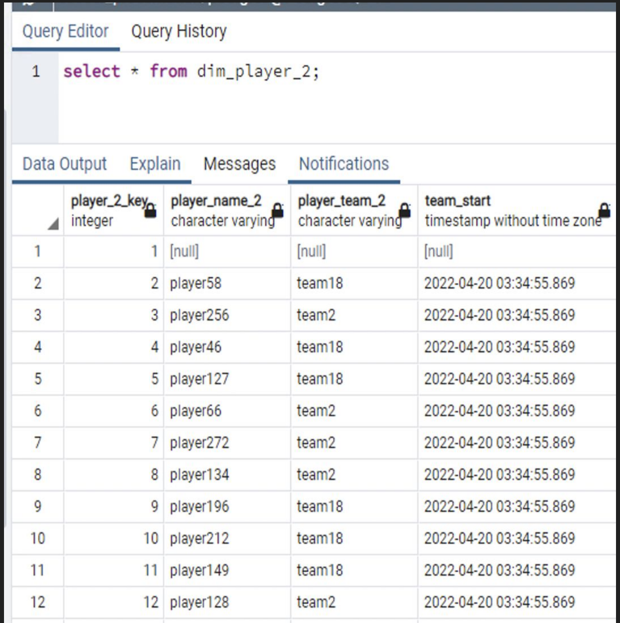
	time_key integer	seconds_from_half integer	seconds_from_start integer	period integer	period_desc character varying	minute integer
1	1	0	0	1	First Half	0
2	2	2	2	1	First Half	1
3	3	3	3	1	First Half	1
4	4	5	5	1	First Half	1
5	5	6	6	1	First Half	1
6	6	9	9	1	First Half	1
7	7	11	11	1	First Half	1
8	8	14	14	1	First Half	1
9	9	15	15	1	First Half	1
10	10	18	18	1	First Half	1
11	11	19	19	1	First Half	1

Sample Data



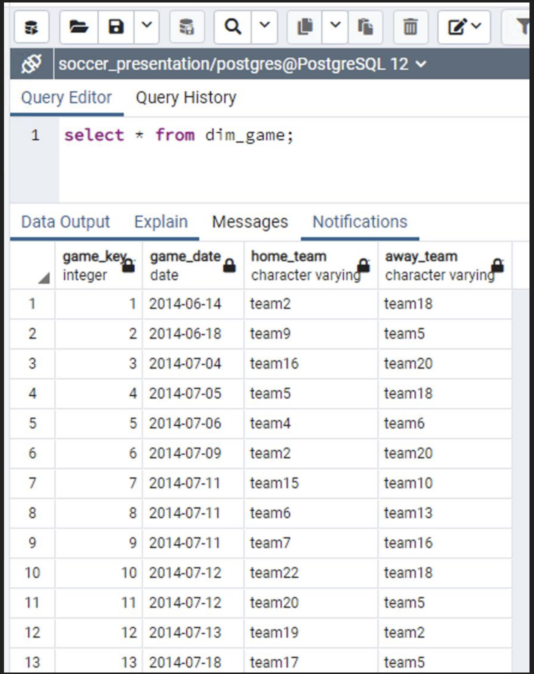
```sql
select * from dim_player_1;
```

	player_1_key integer	player_name_1 character varying	player_team_1 character varying	team_start timestamp without time zone
1	1	player58	team18	2022-04-20 03:16:33.193
2	2	player240	team18	2022-04-20 03:16:33.193
3	3	player246	team18	2022-04-20 03:16:33.193
4	4	player46	team18	2022-04-20 03:16:33.193
5	5	player60	team18	2022-04-20 03:16:33.193
6	6	player283	team18	2022-04-20 03:16:33.193
7	7	player158	team2	2022-04-20 03:16:33.193
8	8	player204	team2	2022-04-20 03:16:33.193
9	9	player256	team2	2022-04-20 03:16:33.193
10	10	player134	team2	2022-04-20 03:16:33.193
11	11	player127	team18	2022-04-20 03:16:33.193
12	12	player233	NULL	2022-04-20 03:16:33.193
13	13	player200	NULL	2022-04-20 03:16:33.193



```sql
select * from dim_player_2;
```

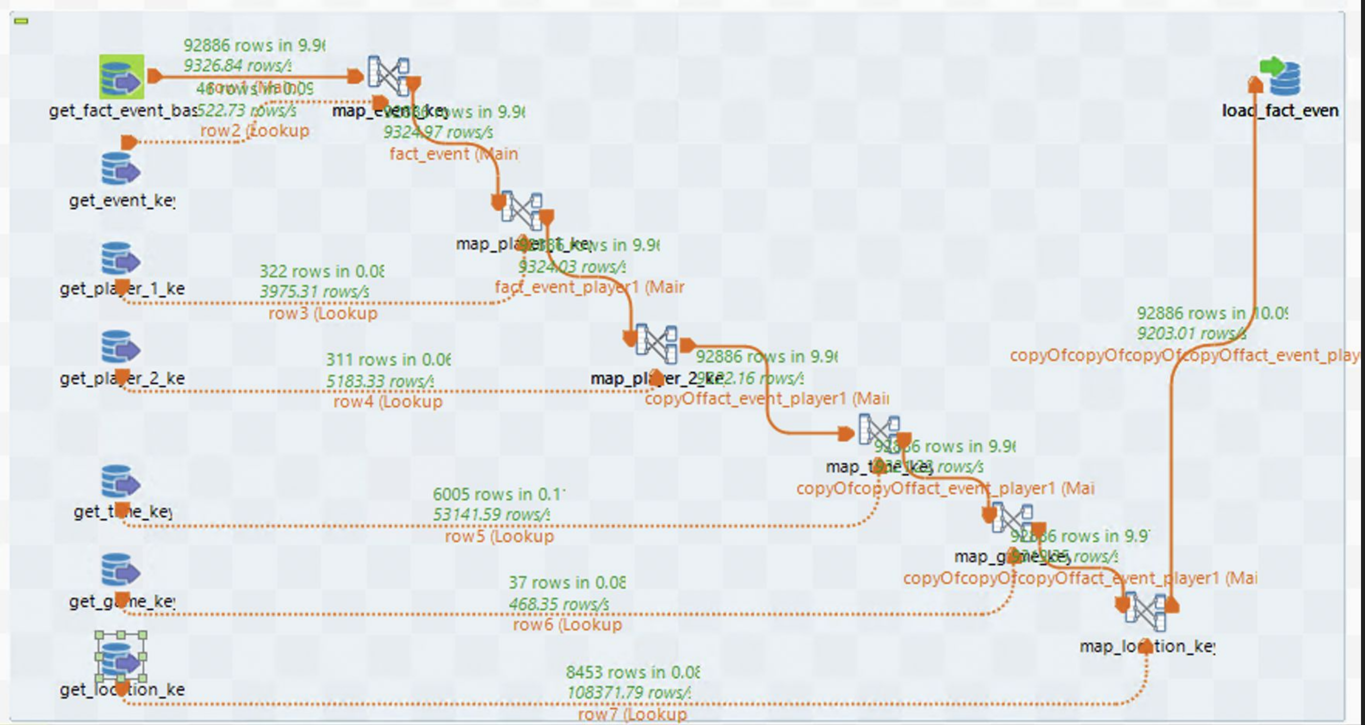
	player_2_key integer	player_name_2 character varying	player_team_2 character varying	team_start timestamp without time zone
1	1	[null]	[null]	[null]
2	2	player58	team18	2022-04-20 03:34:55.869
3	3	player256	team2	2022-04-20 03:34:55.869
4	4	player46	team18	2022-04-20 03:34:55.869
5	5	player127	team18	2022-04-20 03:34:55.869
6	6	player66	team2	2022-04-20 03:34:55.869
7	7	player272	team2	2022-04-20 03:34:55.869
8	8	player134	team2	2022-04-20 03:34:55.869
9	9	player196	team18	2022-04-20 03:34:55.869
10	10	player212	team18	2022-04-20 03:34:55.869
11	11	player149	team18	2022-04-20 03:34:55.869
12	12	player128	team2	2022-04-20 03:34:55.869



```sql
select * from dim_game;
```

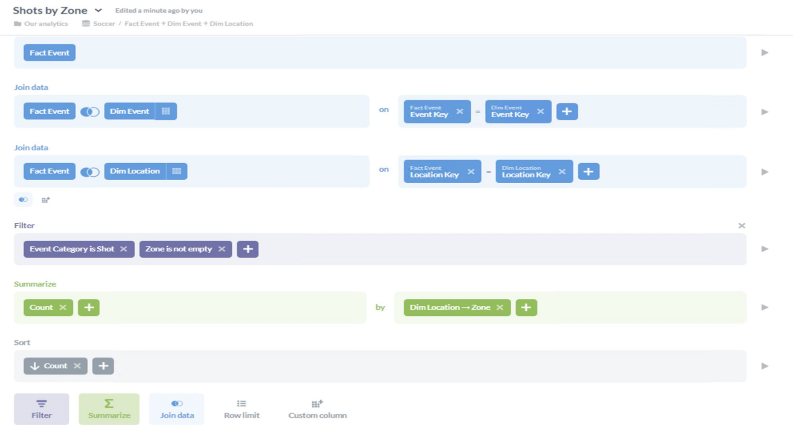
	game_key integer	game_date date	home_team character varying	away_team character varying
1	1	2014-06-14	team2	team18
2	2	2014-06-18	team9	team5
3	3	2014-07-04	team16	team20
4	4	2014-07-05	team5	team18
5	5	2014-07-06	team4	team6
6	6	2014-07-09	team2	team20
7	7	2014-07-11	team15	team10
8	8	2014-07-11	team6	team13
9	9	2014-07-11	team7	team16
10	10	2014-07-12	team22	team18
11	11	2014-07-12	team20	team5
12	12	2014-07-13	team19	team2
13	13	2014-07-18	team17	team5

Fact Table Creation



Metabase Set Up

- Allows for drilling filters
- Easy graphics on dashboard
- Updates when new data enters



Shots by Zone — Edited a minute ago by you

Our analytics · Soccer / Fact Event + Dim Event + Dim Location

Fact Event

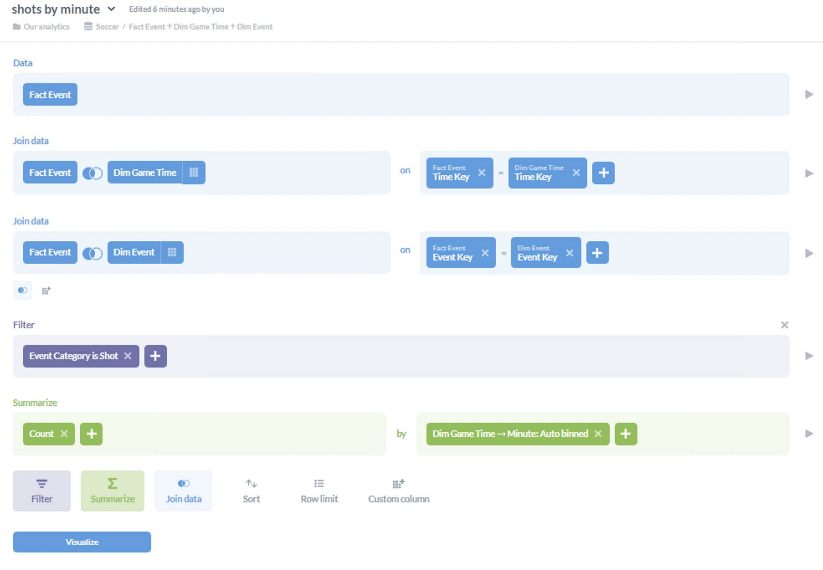
Join data: Fact Event ⟷ Dim Event on Fact Event Event Key = Dim Event Event Key

Join data: Fact Event ⟷ Dim Location on Fact Event Location Key = Dim Location Location Key

Filter: Event Category is Shot · Zone is not empty

Summarize: Count by Dim Location → Zone

Sort: ↓ Count



shots by minute — Edited 6 minutes ago by you

Our analytics · Soccer / Fact Event + Dim Game Time + Dim Event

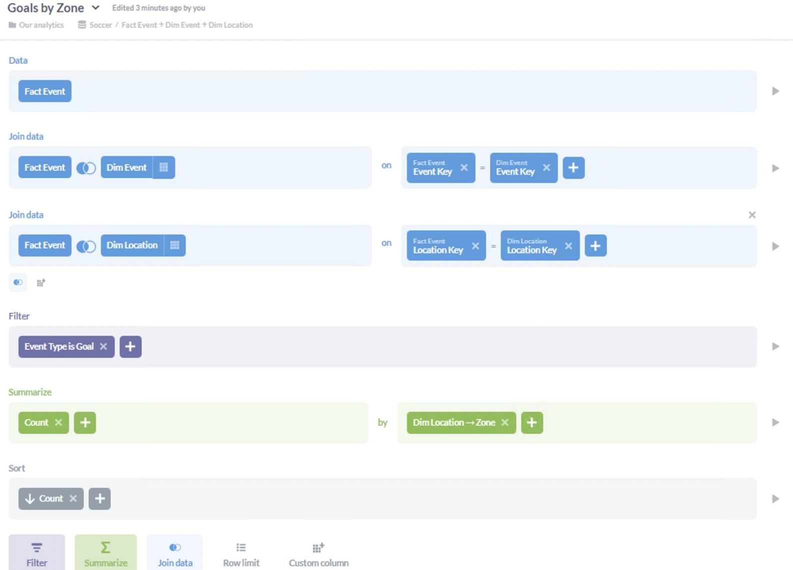
Data: Fact Event

Join data: Fact Event ⟷ Dim Game Time on Fact Event Time Key = Dim Game Time Time Key

Join data: Fact Event ⟷ Dim Event on Fact Event Event Key = Dim Event Event Key

Filter: Event Category is Shot

Summarize: Count by Dim Game Time → Minute: Auto binned



Goals by Zone — Edited 3 minutes ago by you

Our analytics · Soccer / Fact Event + Dim Event + Dim Location

Data: Fact Event

Join data: Fact Event ⟷ Dim Event on Fact Event Event Key = Dim Event Event Key

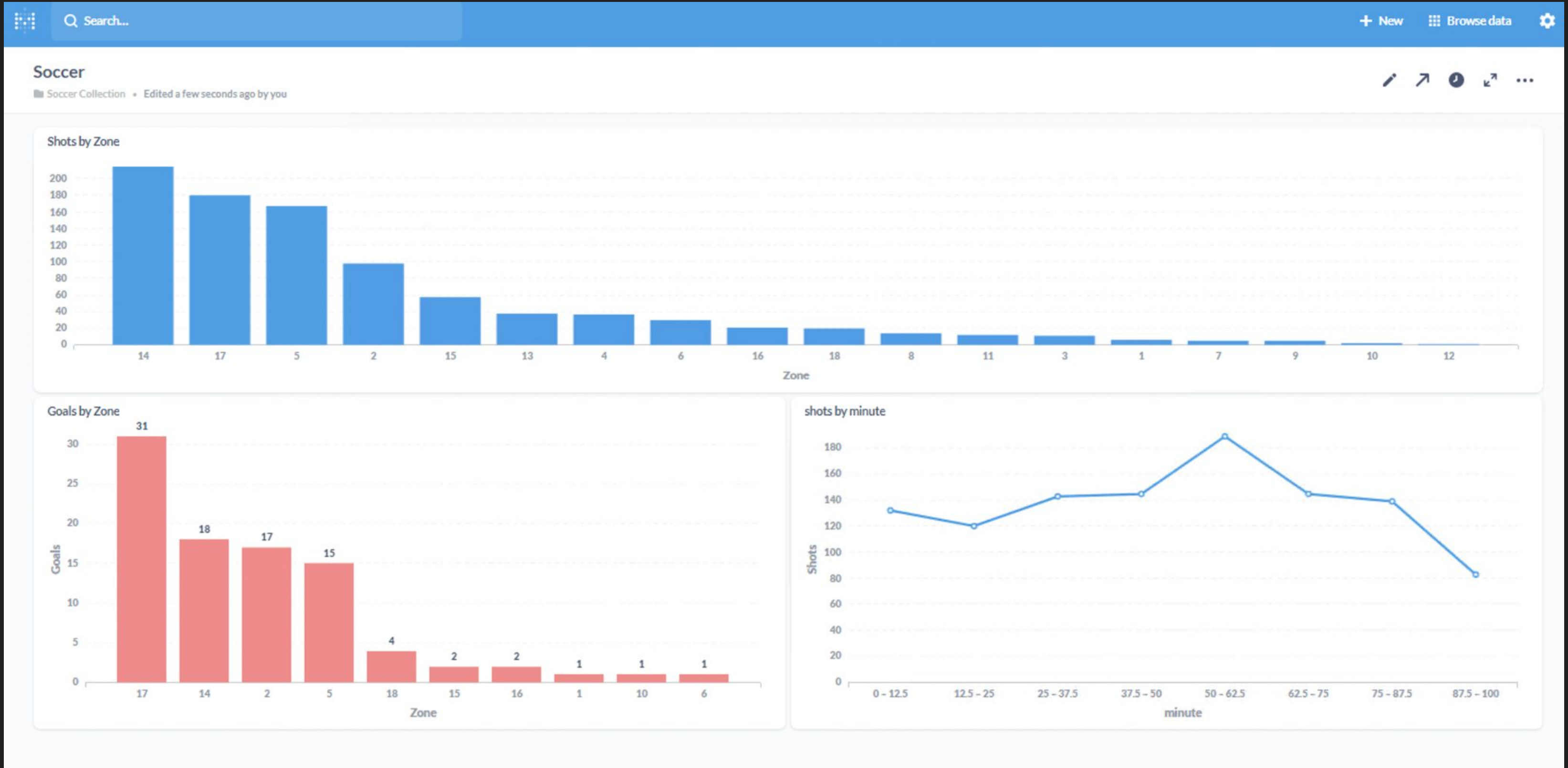
Join data: Fact Event ⟷ Dim Location on Fact Event Location Key = Dim Location Location Key

Filter: Event Type is Goal

Summarize: Count by Dim Location → Zone

Sort: ↓ Count

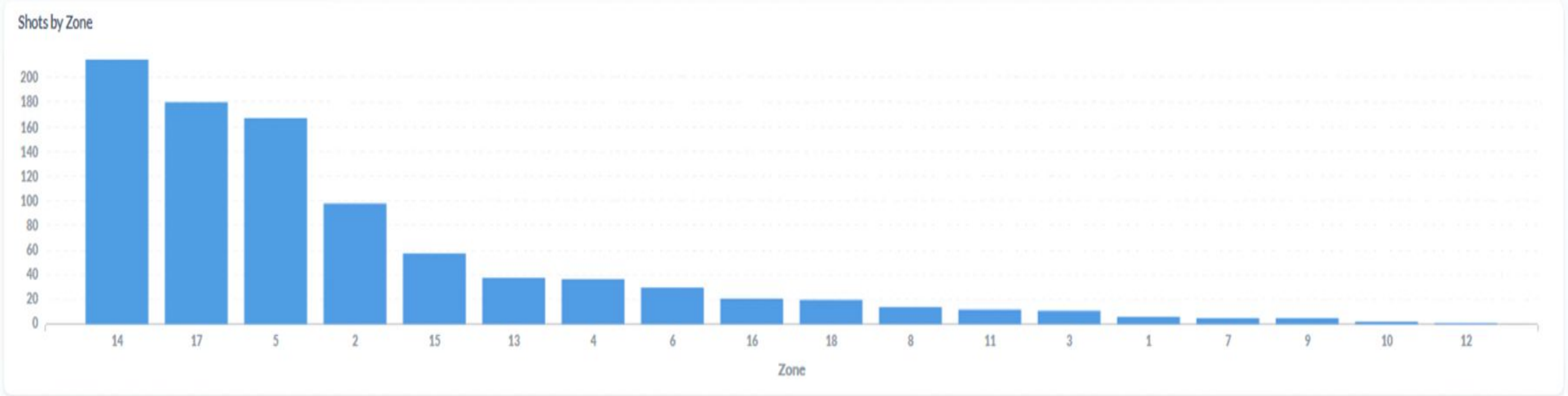
Metabase Dashboard



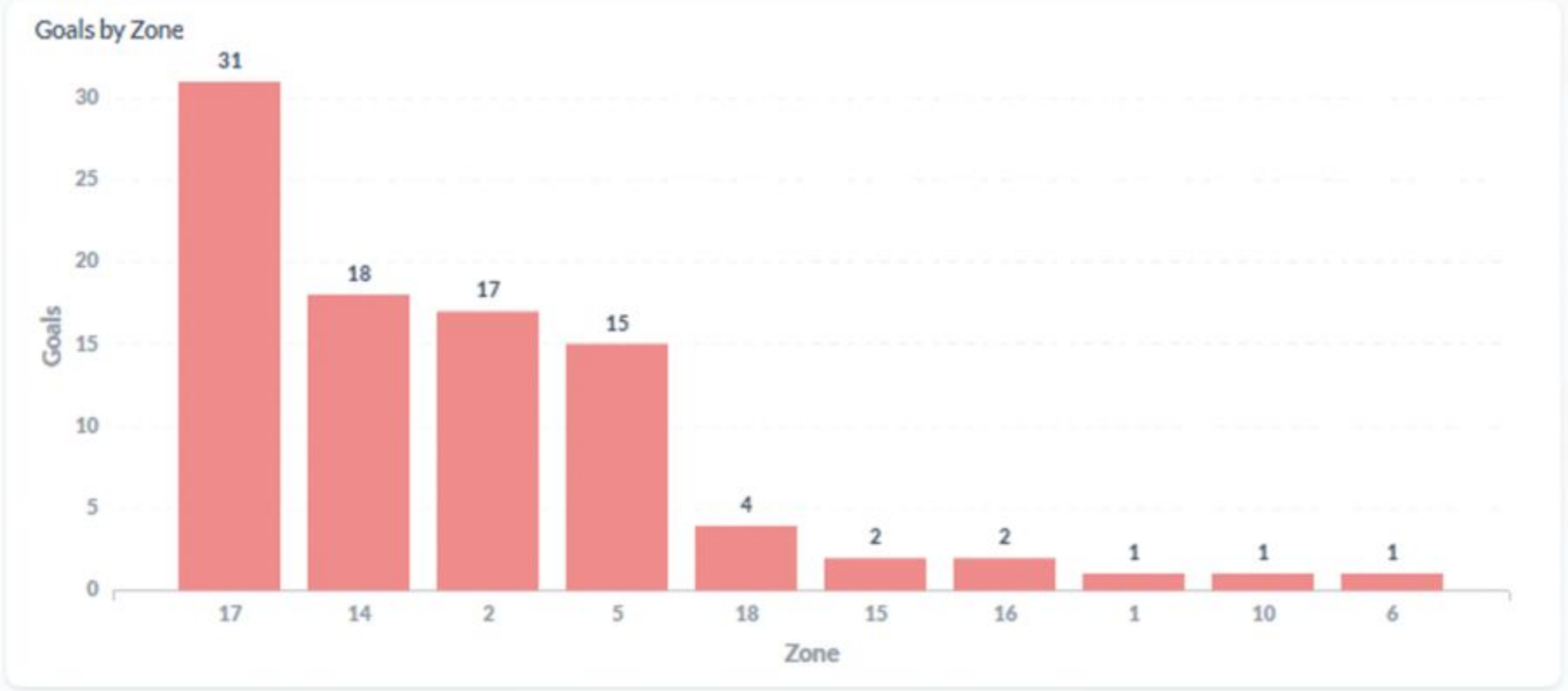
Soccer

Soccer Collection • Edited a few seconds ago by you

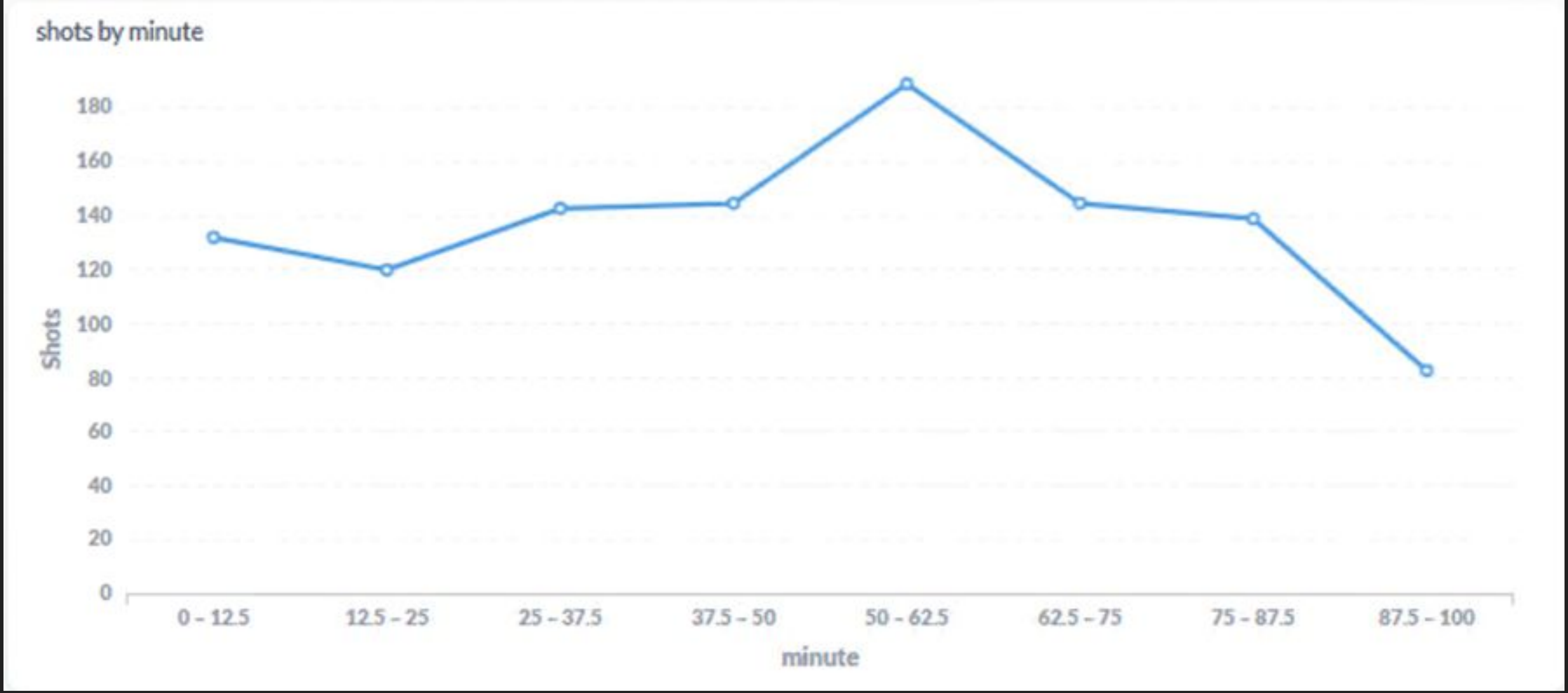
Shots by Zone



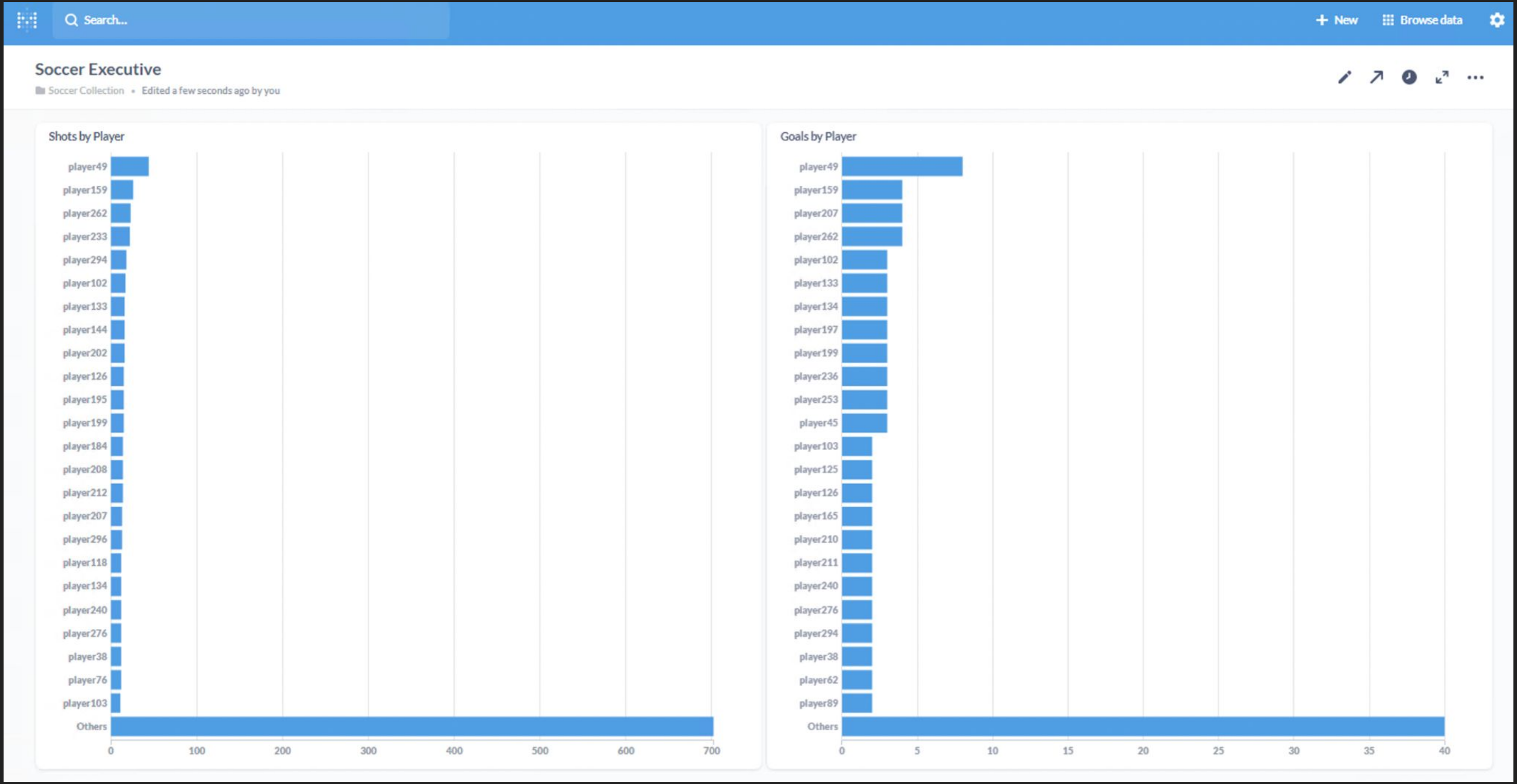
Goals by Zone



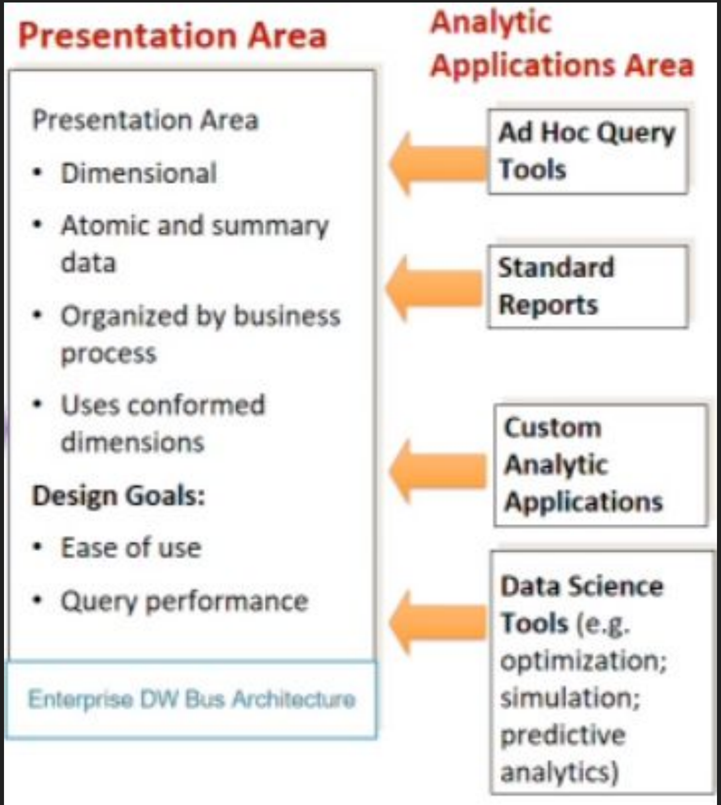
Shots by Minute



Metabase Executive Dashboard



Analytics Ease



Presentation Area

Presentation Area

- Dimensional
- Atomic and summary data
- Organized by business process
- Uses conformed dimensions

Design Goals:

- Ease of use
- Query performance

Enterprise DW Bus Architecture

Analytic Applications Area

Ad Hoc Query Tools

Standard Reports

Custom Analytic Applications

Data Science Tools (e.g. optimization; simulation; predictive analytics)

Future Objectives Requirements

- Where do key passes (passed that lead to shots) come from?
- Where do assists come from?

Conclusion

The business can now:

- Easily load data into a uniform system
- Provide consistent data to analysts
- Manage permissions for specific groups
- Create queries at the speed of thought
- Implement dashboards to easily see how data trends change over time

Presentation Video

https://youtu.be/lwbl8Ni4TPI