MADden

Morgan Bauer 9890-4838
Christan Grant 8143-3970
Joir-dan Gumbs 6148-9357

2011-09-22

1 What are we doing?

We plan to provide a web-based interface for interacting with sports data.

1.1 What are the data sources?

The data we are using is scraped ESPN NFL data, NFL related tweets, team blogs, etc. The scraped ESPN NFL data contains the play-by-play info for every NFL game, player statistics, team statistics, and division blogs.

1.2 The data product

Visualizations of:

1 Correlations between game actions, such as:

A Outcome of coin toss for home/away team

B Probability of making field goal based on stadium (by distance of attempt)

C Fan sentiment with game outcome.

2 Summary of team performance

3 Tweet density for every game

4 ... and other things yet to be discovered

1.3 What is the piece of framework we aim to build?

An extensible, interactive, query-driven framework for sports intelligence. We can run complex queries over heterogeneous data. For example: Given a set of Games G where $G.hometeam$ and $G.awayteam$. and a set of tweets T where for $t \in T$, we can extract: sentiment information $s(t)$ and teamnames $teamnames(t)$.

```
SELECT hometeam, AVG(s(t))
FROM T, G
WHERE hometeam = teamnames(t)
GROUP BY hometeam
```

This result of the this query is a list of the home team sentiment for a selection of games. We could also set a range for the time in the tweet. Notice we had perform both sentiment analysis, information extraction (finding team names) and a poor mans entity-resolution in resolving the team names. This is an example of the type of queries we which to support.

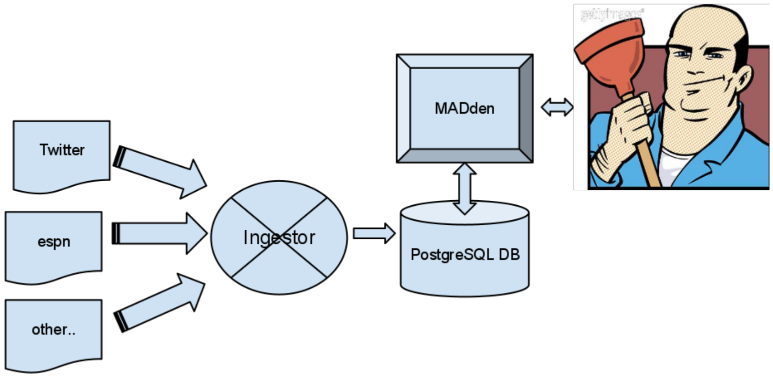

Figure 1: System Architecture

The 1 describes the architecture of our product. We have an Ingestors that pull structured data in from different sources. The ingestor is generic and can be created to adapt to different type of data. The MADden interface takes queries performs client-side logic to all the user to view the results of their queries.

2 What is related work and state-of-the-art

NFL.com, ESPN.com, Yahoo Sports all do sports statistics. Polaris, IBM ManyEyes, Tableau all have done visualization tools. MADLib is what we are basing our project on.

3 Why is it an important project for Data Science?

We are using all tools of data science. Scraping dirty data and turning it into clean information, and then presenting the information in an easy to digest form.

4 What are the novelties of the project? What is the end goal?

A system that allows for tinkering with sports data for gaining insight of team/player performance and fan sentiment.

5 What is the measure of success?

When we are able to perform queries for one user correlating the different datasets, and it ”makes sense.”