The problem that I am solving is helping the New England Patriots determine statistical insights from a large dataset including all 2014 NFL plays. Typical NFL stats include a quarterback's completion rate and total yards, but there are more patterns in the data that could help the Patriots and are only accessible to a data scientist. I imagine that I am hired as a consultant by the Patriots to work with their NFL plays dataset and determine values such as Tom Brady's completion rate to each of his top receivers. I need a database system that is fast, can manage a large amount of data (14+ MB), and provides the capability to do complex queries.

The technology that I utilize to solve my problem and also explain more about in this report is VoltDB. VoltDB database is a new system that runs a database from memory rather than disk. It processes more quickly than disk-based systems and is the first database designed to operate solely on memory. It gaining popularity and customers include well-known enterprises such as HP and Nokia. I downloaded the enterprise trial version onto my Mac and immediately noticed the advantages of memory-based data. Each query or process includes a stop-watch illustrating how quickly the database functions such as loading the NFL dataset in only a few seconds. VoltDB is also user-friendly and designed for MySQL-type users to convert to VoltDB. It features an SQL executable file sqlcmd to access the database with .sql files and similarly has a local-host console and Java API as well. My project begins with directing a .sql file to the sqlcmd executable and then does the data processing through a Java program and Java API.

Overall, my report shows my solution from installation to graphically displaying the results in R.

There are many benefits and advantages to using VoltDB and if you are interested in new generation databases, I recommend downloading VoltDB and trying the steps I outline. VoltDB is a cutting edge database, has recently acquired $23+ million in funding, and

represents the evolution happening in this field and the high quality products available to all users.

Description of Technology

VoltDB is an in memory database that is built for speed. Most databases use disk storage making reads and writes slower and more computationally expensive. Memory sizes are increasing rapidly and VoltDB utilizes this much faster storage to run a database. It is a complete relational database management system (RDBMS) and even supports SQL queries which lessens the learning curve. It is scalable and massively parallel. The makers of VoltDB advertise that it is highly available for datacenter-type use. It also supports enterprise-level functionality such as redundancy and fault tolerance to ensure that data is maintained.

For a first time user, a VoltDB database can be running on a laptop with a single download and only a few configuration steps. Results from commands including reading and processing data are followed with the time it took complete. VoltDB is similar to the most common databases making it easy to learn as opposed to Redis or NoSQL where new storage and access techniques must be learned. Overall, it is a user-friendly database that utilizes memory and offers speed that other disk-based systems cannot match.

Data set

My data set is NFL play-by-play data from 2014. All plays during the NFL season are extracted into tuples in a table. There are over 60 columns of data per row including the offensive and defensive teams, passer, runner, yards gain, and much more. The entire dataset is approximately 14 megabytes and 32,000 rows corresponding to the number of plays in the NFL season. The dataset is from spreadsheet-sports.com and my tests confirm it is complete by calculating aggregated player statistics. The benefit of this dataset is that NFL statistics can be

determined that are not usually available. Sports websites, including NFL.com, feature players' total passing yards but not other statistics like a quarterback's completion rates to each of his top receivers.

Describe your problem statement

I imagine that I am a data analytics consultant helping the Patriots and specifically Tom Brady gain insights from the NFL play-by-play data. My problem is that the Patriots organization has hired me to work with the dataset and to find the less known statistics that can help them win more games. I need to load these 15 MB into a database and perform queries that are beyond the scope of R. The Patriots want to know statistics like total throw attempts to Tom Brady's top receivers and the completion rate to each one. They want to know how many deep balls have been thrown to each of the receivers. Also, they want to know total and average yards to the player types that Tom Brady throws to including wide receivers, running backs, tight ends and more. Overall my challenge is to get the dataset into a fast database, navigate the data with complex queries, and use an enterprise-level system that the Patriots can use in the long-term.

Installation and configuration steps

// Visit the following URL and download the right version of VoltDB for the system. This includes
// giving VoltDB your name and business and confirming your email.
// http://learn.voltdb.com/DLSoftwareDownload.html

// Double click to untar the file **voltdb-ent-6.2.1.tar.gz**.

// Put the untarred file on the Desktop or other convenient location. Change directories into the
// bin folder within the file.
Ryans-MacBook-Pro:Desktop Ryan$ cd voltdb-ent-6.2.1/bin/

// The voltdb create command uses the executable voltdb to make a new database. --force is
// optional but will remove and replace an already running database if needed.
Ryans-MacBook-Pro:Desktop Ryan$ ./voltdb create --force

```
Initializing VoltDB...

 _    __      ____  ____  ____
| |  / /___  / / /_/ __ \/ __ )
| | / / __ \/ / __/ / / / __  |
| |/ / /_/ / / /_/ /_/ / /_/ /
|___/\____/_/\__/_____/_____/

---------------------------------

Build: 6.2.1 voltdb-6.2.1-0-g41a17a9-local Enterprise Edition
Connecting to VoltDB cluster as the leader...
Host id of this node is: 0
Starting VoltDB with trial license. License expires on Jun 13, 2016.
Initializing the database and command logs. This may take a moment...
WARN: This is not a highly available cluster. K-Safety is set to 0.
Server completed initialization.
```

// Once the database is running open in a new terminal and let the server continue
// running in the previous window.
// Now change directories into voltdb-ent-6.2.1/bin/. Run the executable sqlcmd which
// performs SQL commands on the database. A table must be made to hold the data
// before it can be loaded. The file makeTable.sql contains that command.
Ryans-MacBook-Pro:bin Ryan$ ./sqlcmd < makeTable.sql
CREATE TABLE nfl(
index VARCHAR,
Date INT,
Tm VARCHAR,
Opp VARCHAR,
Quarter INT,
Time FLOAT,
Down INT,
ToGo INT,
Side_of_Field VARCHAR,
Yard_Marker INT,
…

// Next the csvloader executable is used to load the NFL play-by-play data into the VoltDB
// database. The data can be downloaded from https://www.spreadsheet-sports.com or the
// dataset copy with this report can be used.
Ryans-MacBook-Pro:bin Ryan$ ./csvloader nfl -f 2014NFLPlay-by-Play_Data.csv --skip=1
Read 31943 rows from file and successfully inserted 31943 rows (final)
Elapsed time: 13.648 seconds
Invalid row file: /Users/Ryan/Desktop/voltdb-ent-6.2.1/bin/csvloader_NFL_insert_invalidrows.csv
Log file: /Users/Ryan/Desktop/voltdb-ent-6.2.1/bin/csvloader_NFL_insert_log.log
Report file: /Users/Ryan/Desktop/voltdb-ent-6.2.1/bin/csvloader_NFL_insert_report.log

// Place the file Program.java onto the Desktop or current location of the voltdb folder.
// Compile the program with the following javac command.
Ryans-MacBook-Pro:Desktop Ryan$ javac -cp .:./voltdb-ent-6.2.1/voltdb/* Program.java

// Run the newly compiled program. The class path for the java command should include
// the jar files into the voltdb subfolder along with the directory for Program.
// Program runs the queries and displays the results with formatted printing. It also
// creates CSVs for the first three queries which will be displayed with R.
Ryans-MacBook-Pro:Desktop Ryan$ java -cp .:./voltdb-ent-6.2.1/voltdb/* Program

TARGETED_RECEIVER	ATTEMPTS	RATE
Julian Edelman	133	0.684210526315
Rob Gronkowski	130	0.623076923076
Brandon LaFell	113	0.619469026548
Shane Vereen	77	0.688311688311
Danny Amendola	33	0.636363636363
Tim Wright	30	0.800000000000
Brian Tyms	11	0.454545454545
Kenbrell Thompkins	11	0.545454545454
NULL	9	0.000000000000
James Develin	8	0.750000000000
Stevan Ridley	5	0.800000000000
Brandon Bolden	5	0.400000000000
Michael Hoomanawanui	5	0.600000000000
Aaron Dobson	5	0.600000000000
LeGarrette Blount	4	1.000000000000
Jonas Gray	3	0.333333333333
Steve Maneri	1	0.000000000000

TARGETED_RECEIVER	PLAYS
Rob Gronkowski	31
Brandon LaFell	19
Julian Edelman	12
Shane Vereen	7
Tim Wright	5
Danny Amendola	2
Brian Tyms	2
Kenbrell Thompkins	1
Aaron Dobson	1
Michael Hoomanawanui	1

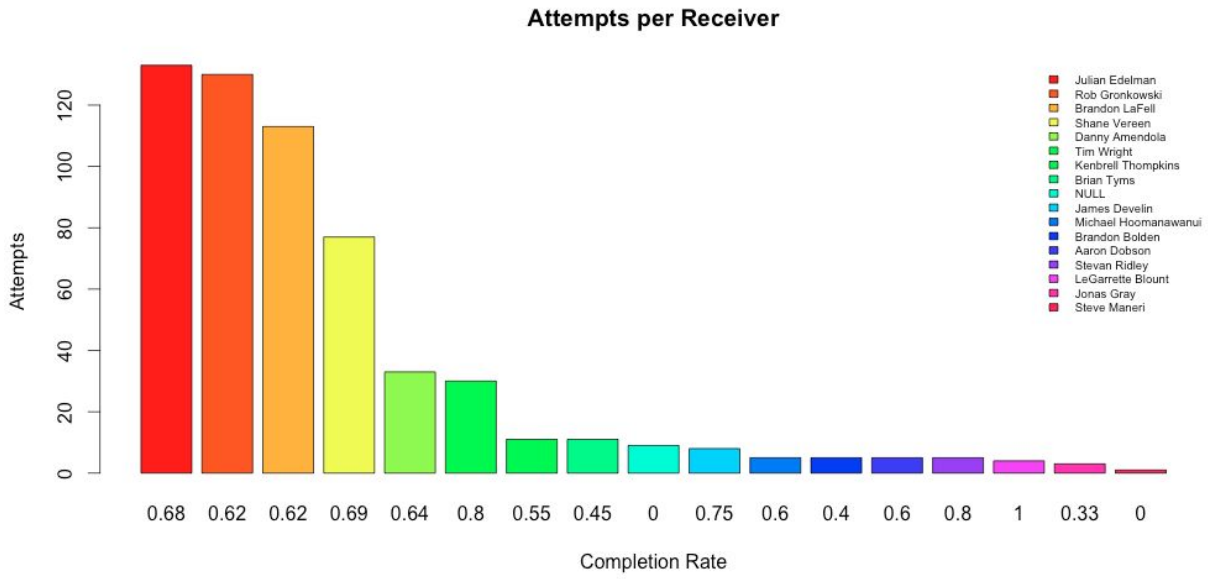
Note: The output screenshots are continued on the next page.

PLAYMAKER_POSITION	TOTALYARDS	AVGYARDS
NULL	82	3.904761904761
RB	543	5.323529411764
TE	1382	8.375757575757
WR	2102	7.125423728813

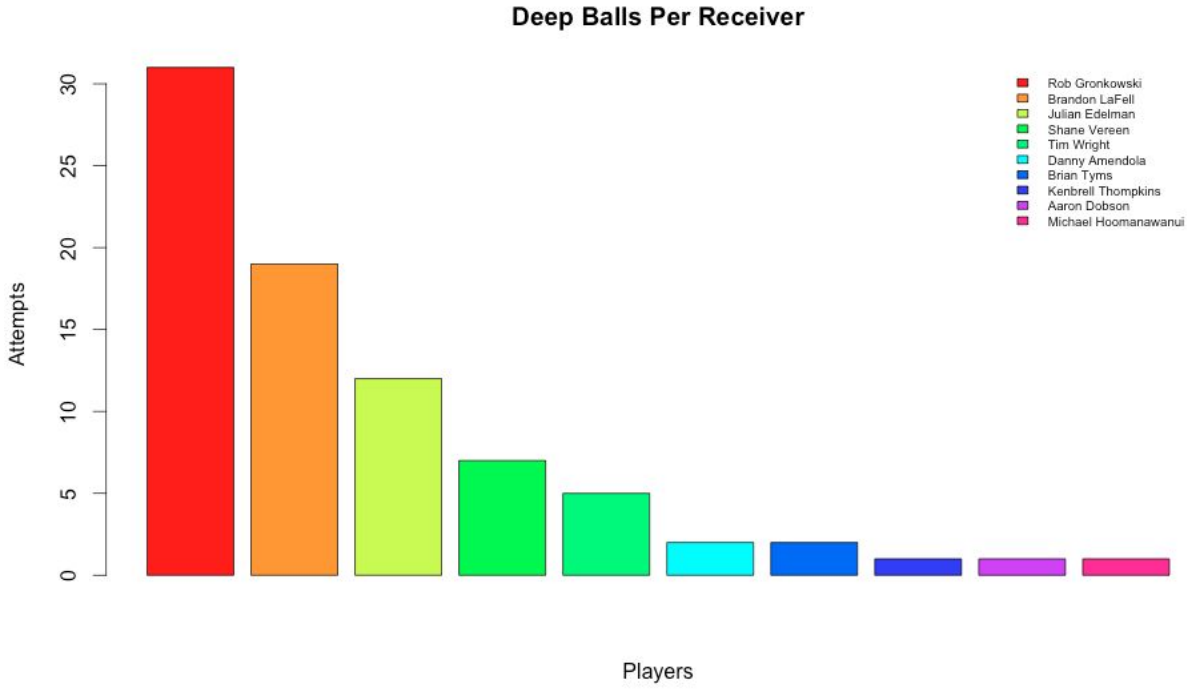
OPP	DEEP_BALLS	RATE
Dolphins	25	0.240000000000
Bills	10	0.500000000000
Broncos	9	0.444444444444
Jets	9	0.555555555555
Bengals	8	0.625000000000
Colts	7	0.285714285714
Packers	7	0.285714285714
Lions	5	0.400000000000
Chiefs	5	0.200000000000
Chargers	4	0.250000000000
Bears	3	1.000000000000
Raiders	3	0.333333333333
Vikings	2	0.000000000000

OPP	THIRD_DOWNS	YARDAGE
Lions	4	11.000000000000
Broncos	5	12.800000000000
Bears	6	12.500000000000
Jets	7	10.857142857142
Chargers	3	12.333333333333
Bills	7	14.571428571428
Bengals	4	13.750000000000
Colts	5	16.800000000000
Chiefs	2	21.000000000000
Raiders	5	12.200000000000
Packers	2	10.500000000000
Vikings	3	20.000000000000
Dolphins	5	14.200000000000

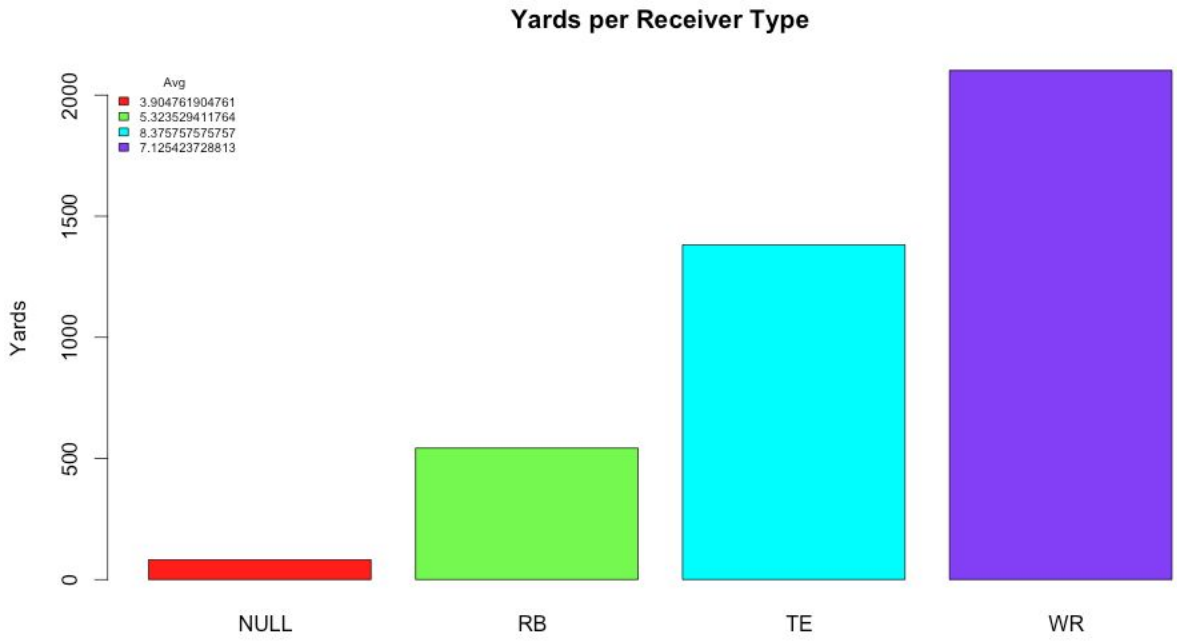
```
// Start RStudio and set the working directory to the location of the voltdb folder, Desktop in this
// case. Run the sections of the R script separately to see each of the visualizations. R should
// be loading the CSV files that were produced in the previous step.
> setwd("/Users/Ryan/Desktop")
> ### 1
> results = read.csv("query1.csv")
>
> barplot(results$attempts, col=rainbow(17), ylab="Attempts", main="Attempts per Receiver",
names.arg = round(as.numeric(results$rate), digits=2), xlab="Completion Rate")
>
> legend("topright", as.character(results$player), cex=0.6, bty="n", fill=rainbow(17));
```




```
> ### 2
>
>
> results = read.csv("query2.csv")
>
> barplot(results$attempts, col=rainbow(10), xlab="Players", ylab="Attempts", main="Deep Balls Per Receiver")
>
> legend("topright", as.character(results$player), cex=0.6, bty="n", fill=rainbow(10));
```




```
> ### 3
>
> results = read.csv("query3.csv")
>
> barplot(results$yards, col=rainbow(4), ylab="Yards", main="Yards per Receiver Type",
names.arg = as.character(results$player))
>
> legend("topleft", as.character(results$rate), cex=0.6, bty="n", fill=rainbow(4), title="Avg");
```



// Optional step: Lastly the Volt management console, accessible through the URL below,
// gives stats on the memory and speed of the database.

Volt management console:
Localhost:8080

What worked, what did not and why not, and lessons learned

Getting the voltdb database installed and running quickly along with loading and querying the data all worked well. VoltDB is designed to be a compelling and user-friendly option for database users to switch to.

The most challenging part was learning the Java API which is utilized in Program.java mentioned earlier. The python API is less polished and VoltDB was originally designed for Java which led me to chose that language and API.

The python API did not work but still consumed a lot of user time, as I tried to focus on it to complete the necessary tasks. The API is newer and less developed than the Java version. I wanted to use Python since the language has big data features and it was my popular choice on homework assignments. It did not work out though and I switched to Java.

I learned that teaching myself a new database system is not as difficult as I imagined. The companies like VoltDB have extensive user support and documentation since they are trying to get customers to convert their system. It was useful to gain knowledge and experience with a cutting edge database system like VoltDB. It was also a good illustration of the speed of memory, at a time when many laptops are converting to all flash memory.