

Data Analytics using **Yelp Data**

Document name: Final Project Report

Subject: Csc 230

Semester: Spring 2015

Group #5

Members: Nevil Patel, Suraj Ponugoti

Professor: Doan H Nguyen

Contents

Glossary of terms	3
Customer Statement of Requirements.....	4
Goals	4
Problem Statement	4
Proposed Solution Statement	4
Enumerated Functional Requirements.....	7
List of enumerated Nonfunctional Requirements	8
System diagram	9
Functional Requirements Specification	10
Stackholders.....	10
Actors and Goals.....	10
Casual Description	10
Use Case Diagram	12
Identify factors for potential successful business (Figure 1)	12
Improve customer services by identifying futuristic factors (Figure 2)	13
Identify Cultural Trends (Figure 3)	13
Use Case Description.....	14
Traceability Matrix.....	17
Class Diagram and Interface Specification	18
Sequence Diagram	19
User Interface Design and Implementation	20
Testing	23
Project Status	25
Current Status	25
Future Enhancements	25
Lessons Learned	25
Contribution of CSC 230 course to this project	25
References	25

Glossary of terms

Yelp: A website which publishes crowd source reviews to help users and business owners .

Business: A local body listed on yelp like Restaurants, Department Stores, Bars, Home-Local Services, Cafes, Automotive.

Existing business owner: A person who has listed his business on Yelp site and getting views from yelp users.

Future business owner: A person who wants to start new business in future time

User: A person who has registered on yelp who is writing reviews about different business after visting them or a person who is using yelp reviews to choose business.

Analytics: Extract knowledge out of data which can be used by system users to make important decisions which is very difficult just by looking at the data.

Review: It is text written by user after visting business about the over-all experience. I is also a numeric representation (out of 5) to compare it with other business.

Customer Statement of Requirements

Goals

The objective is to design system which will use existing yelp data to provide insightful analytics and help existing business owners, future business owners to make important decisions regarding new business or business expansion.

Problem Statement

40% of world population has internet connection today compared to 1% in 1995. Almost 3 exabytes of data is created per day using internet. Storing huge amount of data and retrieving knowledge out of it is challenging task these days. Yelp is a website which publishes crowd sourced reviews about local businesses (Restaurants, Department Stores, Bars, Home-Local Services, Cafes, Automotive). It provides opportunity to business owners to improve their services and users to choose best business amongst available.

Currently Yelp just lists all user reviews and provide average ratings for particular business. It is very difficult for business owners to go through list of reviews and carve out important information out of it. Also it does not provide advanced analytics to business owners to grow their business and improve their services. Our solution will focus on providing advance analytics from yelp data to help existing business owners, future business owners and users.

Proposed Solution Statement

Here are the main points we are including in our solution:

- We want to enable existing business owners to improve their services using analytics and also they can make important decisions regarding business expansion in new cities, countries.
- We want to enable future entrepreneurs to know what the success chances are if they open particular business in particular city or country. They can also determine which attributes (Wifi, Parking, ambience etc.) play major role when it comes to attracting more customers.
- We want to enable government organizations to figure out few interesting cultural trends using yelp data.

There are total 10 functional requirements. We will use hadoop map reduce algorithm with Java Api to process the large dataset and implement those 10 functional requirements. So there will be 10 different hadoop map reduce programs which will implement these 10 functional requirements. We will also develop a web application which will show results of these programs in integrated manner using 1 user-interface.

Our system will analyze past check-in and review dataset and determine the day of week and during which hours of the day business will be busier. Existing business owners can use this information and take pre cautionary steps to employ more staff so that their services won't be degraded when there is more rush.

Restaurants or bars have normal rush during regular days but on special occasions, they have more rush and they can't cop up with that which turns in to bad customer service. Sports bars can have more rush on major game days and restaurants on festival days or on weekends. Using past analytics data, our system will determine that how many more visitors' particular sports bar can expect on major game days so that they can prior arrangements to handle the rush. Also System will determine that how many more visitors' restaurants with bars can expect on Fridays and Saturdays.

Our system will enable existing and future business owners to decide whether location is important factor for business's success or not. This way future entrepreneurs can determine the best profitable location for business and existing business owners can decide the location for opening branches of their existing business in different areas.

Future entrepreneurs always have this question in mind that where and what kind of business they should open to gain maximum profit. Using our solution, business owners will be able to determine that what kind of business is more profitable and attract more users in particular city or area. They will be able to determine that Restaurants (Mexican, Chinese etc.), Bars (Sports Bars, Dive Bars, and Nightlife), Shopping (Home & Garden, Furniture Stores, Gift etc.) are more favorable to succeed in particular city or country.

Apart from offering regular services, certain businesses lure customers by offering extra set of services. Our system will determine that,

- Coffee shops, restaurants, night bars having wifi attract more visitors than business not having wifi?
- Business with more parking slots attract more visitors than business with less parking slots or no parking slots nearer to business.
- Business which does not accepts Credit Cards are less likely to be preferred by customers or is it equally preferred.

Also our system will determine few cultural trends like,

- When Germans, UK, USA, Canadian people prefer to eat?
- What cuisines people like in Germany, UK, USA, Canada?

Enumerated Functional Requirements

Identifier	Priority	Requirement
REQ-1: Identify Day & Hours	5	System shall determine that on which day of week and during which hours of the day business will be busier.
REQ-2: Determine surge (%) in visitors	5	System shall determine surge in visitors(%) sports bars can expect on game days and How many more visitors restaurants with bars can expect on Fridays and Saturdays.
REQ-3: Determine location factor in success	5	System shall determine that how much of a business's success is really based upon just location.
REQ-4: Favorable type of business in city	5	System shall determine type of business which is more favorable to succeed in particular city .
REQ-5: Favorable type of business in country	4	System shall determine type of business which is more favorable to succeed in particular country .
REQ-6: Wifi factor in business success	4	System shall determine that the coffee shops, restaurants, night bars having wifi attract more visitors.
REQ-7: Parking factor in business success	4	System shall determine that business with more parking slots attract more visitors than business with less parking slots or no parking slots nearer to business.
REQ-8: Credit Card factor in business success	4	System shall determine that business which does not accepts Credit Cards are less likely to be preferred by customers or is it equally preferred.
REQ-9: Cultural Trend 1	3	System shall determine when Germans, UK, USA, Canadian people prefer to eat?
REQ-10: Cultural Trend 2	3	What cuisines people like in US, UK, Canada, Germany?

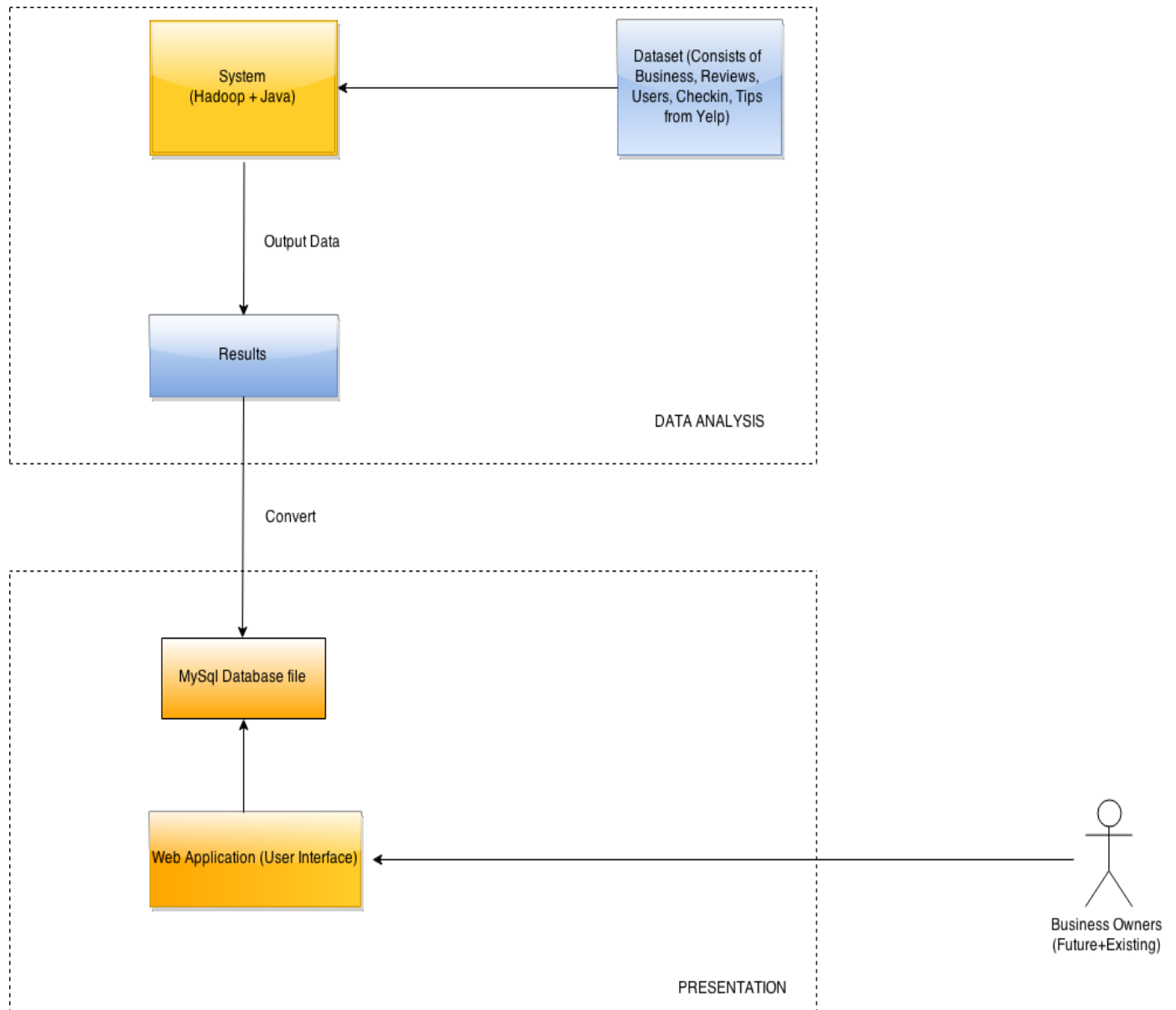
List of enumerated Nonfunctional Requirements

Identifier	Priority	Requirement
REQ-11: Accurate Distance	5	System should accurately measure the business's distance from the prime area of city. (Not more than 0.2 mile difference)
REQ-12: Accurate Prediction	4	System should accurately measure surge of visitors in % unit. (Error rate should be below 1%)

FURPS Table

FURPS (Priority Five)	
Functionality	<ul style="list-style-type: none">• Hadoop map-reduce algorithms are used for processing big data sets.• A web application will display results as text and graph output.
Usability	<ul style="list-style-type: none">• User and business owners will have simple user interface to see analytics.• User can also see graphs when it is difficult to understand text data.
Reliability	<ul style="list-style-type: none">• System discards the data row if it does not provide enough information or if it is corrupt
Performance	<ul style="list-style-type: none">• For better performance, system will perform analytics on 1.32 gb dataset which has enough information to predict or determine various factors.

System diagram



Functional Requirements Specification

Stakeholders

- Existing business owners registered on web
- Future entrepreneurs who wants to open new business
- Government Organizations who wants to identify cultural trends

Actors and Goals

Actors	Goals
Future entrepreneur	He wants to analyze the existing business data provided by yelp so that he can open successful business.
Existing Business Owner	It is registered business owner on yelp along with his business and wants to improve services using system.
Government Organizations	They use the yelp data and system to identify some cultural trends in countries like USA, UK, Germany and Canada.

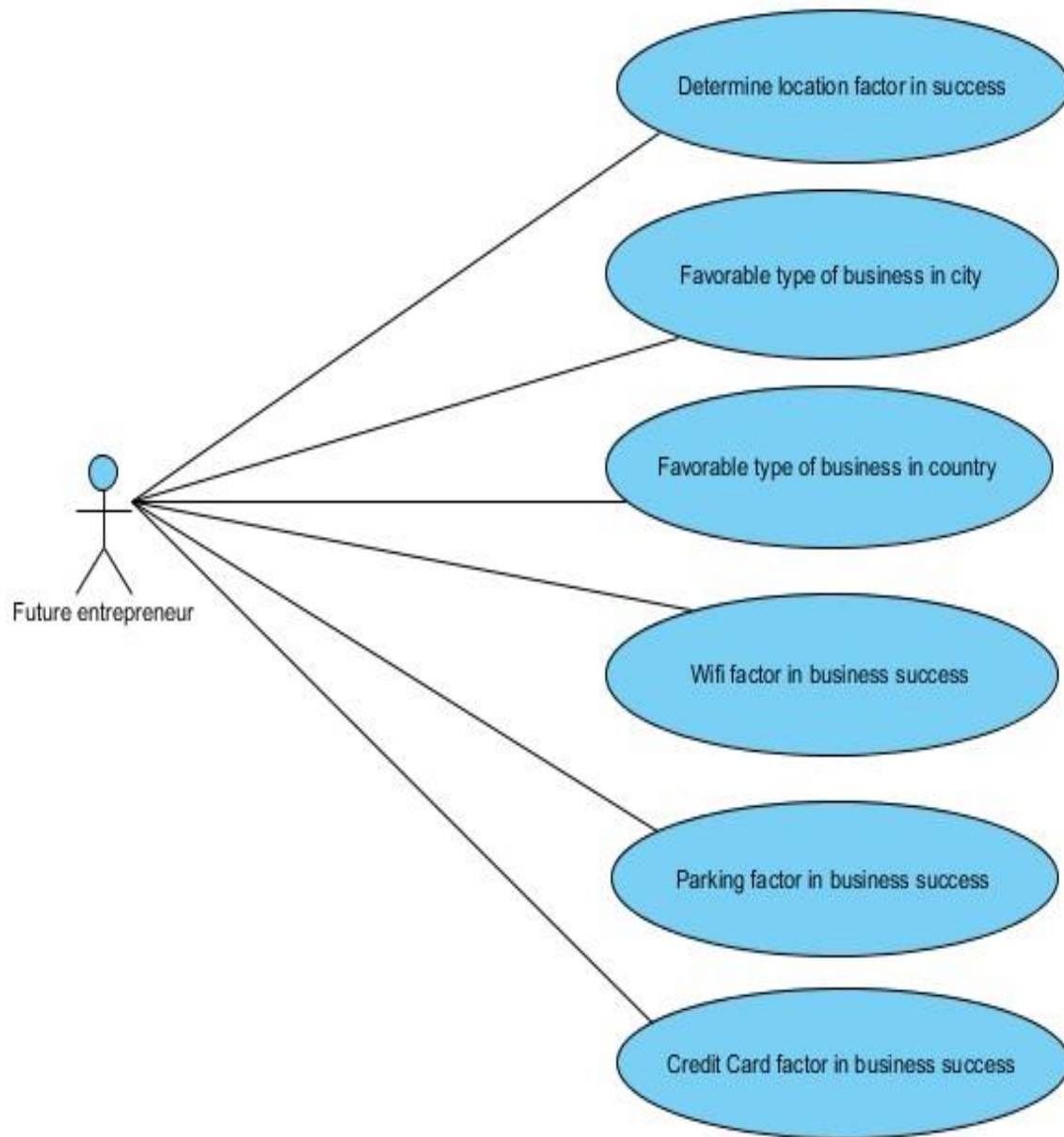
Casual Description

Use Case	Identifier	Description
UC-1	Identify Day & Hours	Determine that on which day of week and during which hours of the day business will be busier.
UC-2	Determine surge (%) in visitors	Determine surge in visitors(%) sports bars or restaurants can expect on game days or on Fridays, Saturdays
UC-3	Determine location factor in success	Determine that how much of a business's success is based upon location.
UC-4	Favorable type of business in city	Determine type of business which is more favorable to succeed in particular city.
UC-5	Favorable type of business in country	Determine type of business which is more favorable to succeed in particular country.
UC-6	Wifi factor in business success	Determine that the coffee shops, restaurants, night bars having wifi attract more visitors or not.
UC-7	Parking factor in business success	Determine that business which does not accepts Credit Cards are less likely to be preferred by customers.

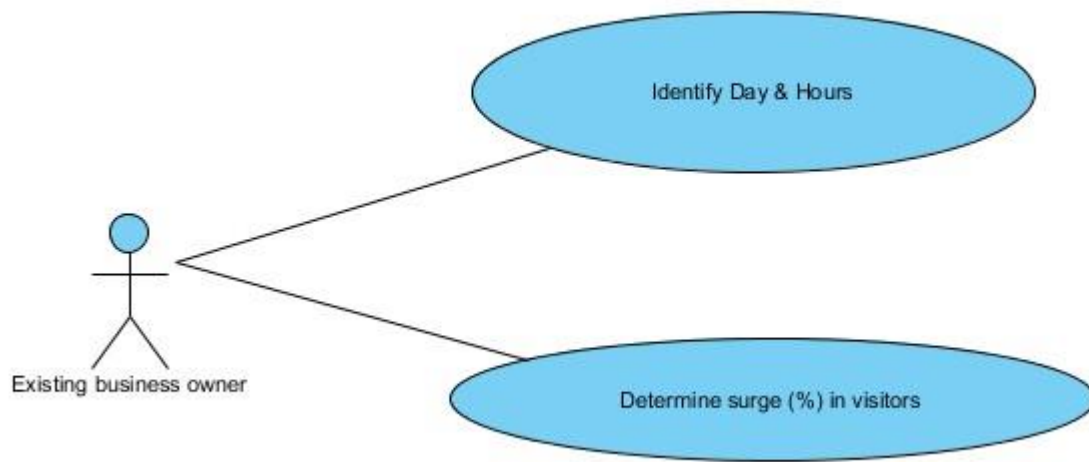
UC-8	Credit Card factor in business success	Determine that business with more parking slots attract more visitors than business with less parking slots or no parking slots nearer to business.
UC-9	Identify Cultural Trends	Determine few cultural trends of people living in USA, UK, Germany, Canada.

Use Case Diagram

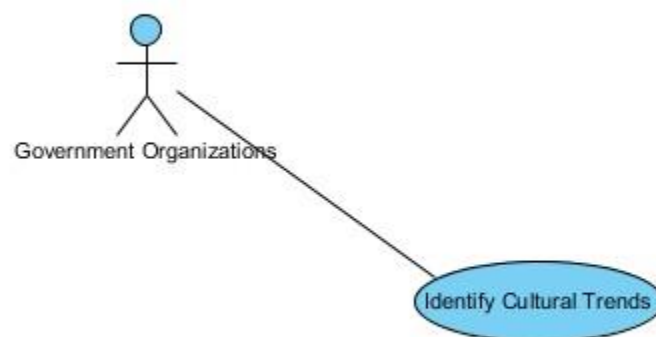
Identify factors for potential successful business (Figure 1)



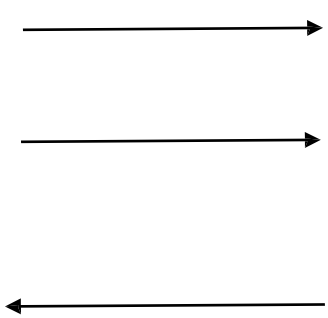
Improve customer services by identifying futuristic factors (Figure 2)



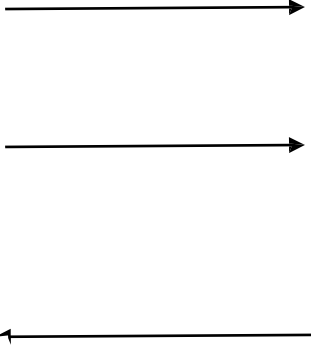
Identify Cultural Trends (Figure 3)

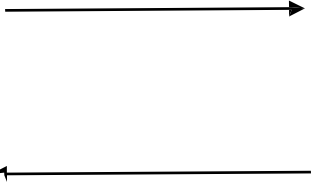


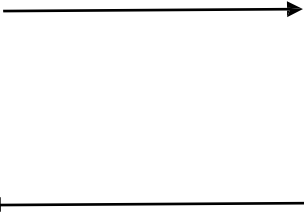
Use Case Description

Use Case 1	Identify Day & Hours
Initiating Actors	Existing business owner registered on yelp
Actor's Goal	Determine that on which day of week and during which hours of the day his business will be busier.
Precondition	Business owner should be registered on yelp along with his business and business should have sufficient information which can be processed by system.
Post condition	System identifies busiest day and hours for any business and returns to user.
Main Success Scenario	<div><pre>graph TD; A[] --> B[]; B --> C[]; C --> D[];</pre></div> <p>(1) Business owner selects 1 of the business he is owing.</p> <p>(2) Business owner selects whether he needs weekly or daily information.</p> <p>(3a) If system does not have enough information, It will show error.</p> <p>(3b) If enough information is available, System will process the information and outputs day and time when business will be busier.</p>

Use Case 4	Favorable type of business in city
Initiating Actors	Future business owner (entrepreneur)
Actor's Goal	Determine type of business which is more favorable to succeed in particular city .
Precondition	Future Business owner who has access to data provided by yelp and logged into system.
Post condition	System determines favorable type of business for city chosen by user.

<p>Main Success Scenario</p>  <pre> graph TD A[] --> B[] B --> C[] C --> A </pre>	<p>(1) User selects the city in which he wants to start new business.</p> <p>(2) User selects type of business he wants to get into. (Restaurants (Mexican, Chinese etc.), Bars (Sports Bars, Dive Bars, Nightlife), Shopping (Home & Garden, Furniture Stores , Gift etc.)</p> <p>(3) System process the existing data for all the cities for the given business type and tells whether given city is favorable for opening that business or not.</p>
--	--

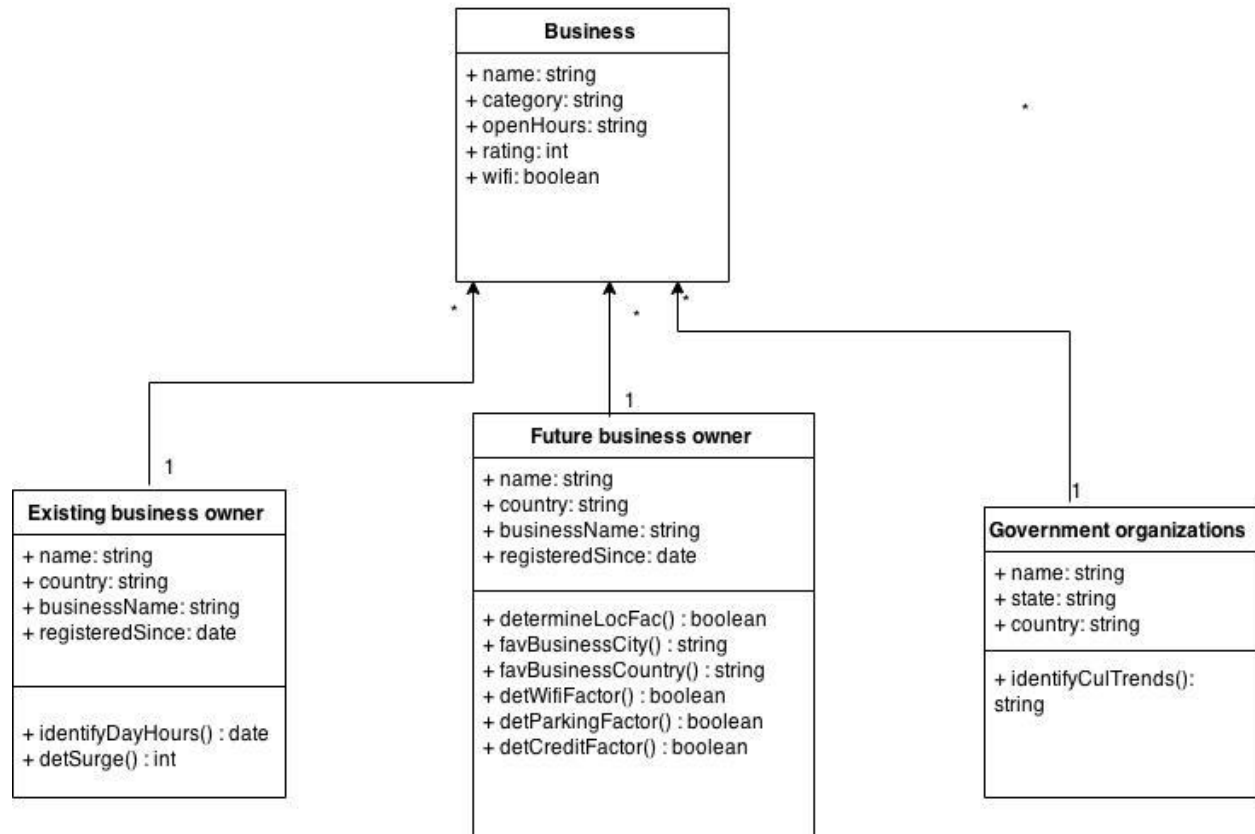
Use Case 2	Determine surge (%) in visitors
Initiating Actors	Existing business owner registered on yelp
Actor's Goal	Determine surge in visitors (%) sports bars can expect on game days and How many more visitors, restaurants with bars can expect on Fridays and Saturdays.
Precondition	Business owner should be registered on yelp along with his business and business should have sufficient information which can be processed by system.
Post condition	System determines surge is visitors for any business. (Unit: %)
<p>Main Success Scenario</p>  <pre> graph TD A[] --> B[] B --> A </pre>	<p>(1) Business owner selects whether he wants to check surge in visitors for Bars or restaurants.</p> <p>(2) System processes data and displays % of more visitors owner can expect on Fridays, Saturdays or on sports days.</p>

Use Case 6	Credit Card factor in business success
Initiating Actors	Future business owner (entrepreneur)
Actor's Goal	Determine that business which does not accepts Credit Cards are less likely to be preferred by customers (in %) or is it equally preferred.
Precondition	Future Business owner who has access to data provided by yelp and logged into system.
Post condition	System determines whether credit card is important factor in business success or not.
Main Success Scenario	<div>  </div> <p>(1) User selects type of business he wants to get into. (Restaurants (Mexican, Chinese etc.), Bars (Sports Bars, Dive Bars, Nightlife), Shopping (Home & Garden, Furniture Stores, Gift etc.)</p> <p>(2) System processes data and tells whether given business type without credit card accepting facility is less</p>

Traceability Matrix

	UC-1	UC-2	UC-3	UC-4	UC-5	UC-6	UC-7	UC-8	UC-9
REQ-1: Identify Day & Hours	X								
REQ-2: Determine surge (%) in visitors		X							
REQ-3: Determine location factor in success			X						
REQ-4: Favorable type of business in city				X					
REQ-5: Favorable type of business in country					X				
REQ-6: Wifi factor in business success						X			
REQ-7: Parking factor in business success							X		
REQ-8: Credit Card factor in business success								X	
REQ-9: Cultural Trend 1									X
REQ-10: Cultural Trend 2									X
REQ-11: Accurate Distance			X						
REQ 12: Accurate Prediction		X				X	X	X	

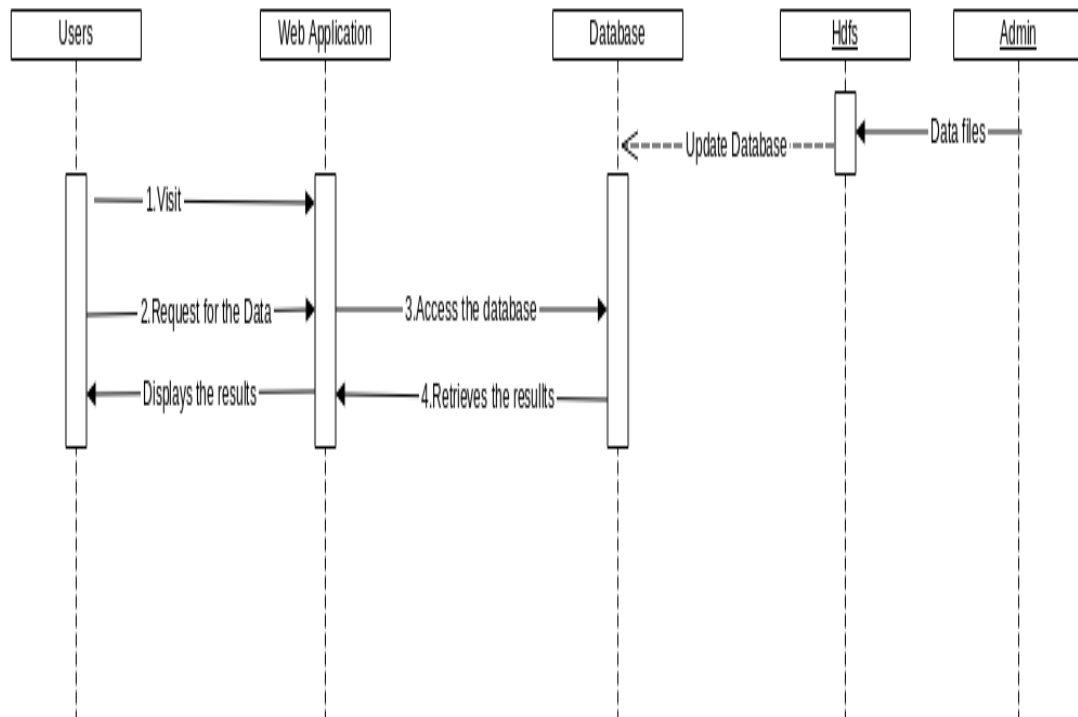
Class Diagram and Interface Specification



Sequence Diagram

For our system, there are common sequence diagram for all use cases. So this is general sequence diagram for all use case.

Users can visit the web application and access the functionalities. Web application gets new data from Hdfs whenever new set of data is available.



User Interface Design and Implementation

(1) Successful type of business in city

Result determines type of business which is more favorable to succeed in **particular city**.

```
hduser@Nevil-Linux:/usr/local/hadoop$ bin/hdfs dfs -cat func4_output/*
15/04/13 16:56:59 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Different cities      Successful types of Restaurants
Edinburgh            Fast Food
Karlsruhe            Italian
Montreal             Italian
Waterloo             Chinese
Pittsburgh           American
Charlotte            American
Urbana               American
Champaign            American
Phoenix              American
Las Vegas            American

Different cities      Successful types of Real Estate Business
Edinburgh            Real Estate Services
Karlsruhe            Real Estate Services
Montreal             Real Estate Services
Waterloo             Real Estate Services
Pittsburgh           Real Estate Services
Charlotte            Property Management
Urbana               Property Management
Champaign            Property Management
Phoenix              Real Estate Services
Las Vegas            Real Estate Services

Different cities      Successful types of Automotives
Edinburgh            Gas & Service Stations
Karlsruhe            Auto Repair
Montreal             Auto Repair
Waterloo             Tires
Pittsburgh           Auto Repair
Charlotte            Auto Repair
Urbana               Auto Repair
```

(2) Successful type of business in country

Result determines type of business which is more favorable to succeed in **particular country**.

```
hduser@Nevil-Linux:/usr/local/hadoop$ bin/hdfs dfs -cat func5_output/*
15/04/13 16:53:05 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable

Different countries      Successful types of Restaurants
U.K.    Fast Food
Germany Italian
Canada  Italian
U.S.    American

Different countries      Successful types of Real Estate Business
U.K.    Real Estate Services
Germany Real Estate Services
Canada  Real Estate Services
U.S.    Real Estate Agents

Different countries      Successful types of Automotives
U.K.    Gas & Service Stations
Germany Auto Repair
Canada  Auto Repair
U.S.    Auto Repair

Different countries      Successful types of Active Life
U.K.    Golf
Germany Bowling
Canada  Amusement Parks
U.S.    Amusement Parks
```

(3) Parking Factor

This functionality determines whether parking factor plays major role in business success or not.

```
hduser@suraj-Ideapad-Z570: /usr/local/hadoop
WI Madison Tenney - Laphan r20tJRPrDnaMplISJ9I9_w true 3.5 14
WI Madison Tenney - Laphan xtW28KmwlgJ8FAPgT60kFw true 4.0 71
WI Madison ThiererRd DgKTPedH5yk552xBUSlgCw false 2.5 13
WI Madison ThiererRd jqo3Ljexof9sA8PhSTwEJA true 4.0 70
WI Madison TreeLn Jcy20ENZ9uTPee2gE6InMg true 4.0 39
WI Madison TreeLn xCoAdsSkgwEm-hzKN25NHA true 2.5 17
WI Madison UnivAv RC_6Br1TzdgWDDIZMAI1A true 3.5 31
WI Madison UniversityAve 2uapF8uioz-DZ2Eq0eeXmQ false 3.5 20
WI Madison UniversityAve JQrF4lHYZkf5IBTVQ8pIw true 3.0 113
WI Madison UniversityAve JxYTEIQqDEXgjHXtfvecAA false 2.5 3
WI Madison UniversityAve N6G0zN3tUw1uZa9F3TK6ZQ false 2.5 53
WI Madison UniversityAve WqAGSp6oSKPheFRjbWkdnQ true 3.0 19
WI Madison UniversityAve YAD6k60BBxgevBkyXgKE4Q false 3.0 5
WI Madison UniversityAve YXfDJLWtchsCccaHMMkyYw true 3.0 70
WI Madison UniversityAve dgq6Y3t81Hq30g-qd4C5Ew false 3.0 28
WI Madison UniversityAve jRRXj-_BhNvmxG8I568uKg true 4.0 34
WI Madison UniversityAve lWtpoG_7K7wmyvggpdIDRQ true 4.0 19
WI Madison UniversityAve omWFwLRqI3VTcM46L8-qaA true 3.5 75
WI Madison UniversityAve ywbedAAD0n5kP-F4dwAdeA false 3.0 4
WI Madison UniversityAvenue lvtZQyk8W-ZRbUHKUpXh6Q false 4.0 47
WI Madison VeronaRd 3Jq5cde5LN9tLPrrouAEFw true 4.0 36
WI Madison VeronaRd oZaCeojhVeBMvcjk_bFDag false 4.0 7
WI Madison Vilas 1s5nBRc-6BqFT_1-ZaVVug true 4.0 63
WI Madison Vilas BfiU8hGgUpNNm7t-ljnmw true 4.0 27
WI Madison Vilas Bgsr91YOV0F48h00UCgcIA true 4.0 75
WI Madison Vilas GMy44IklfrxF-o4TeJf48Q false 4.0 11
WI Madison Vilas J9e170eP0gmkaJR1pEFB3Q true 3.5 44
WI Madison Vilas JeXAH4ViHZpZi7RMC-NQjw true 3.0 50
WI Madison Vilas TsUWz3TMkFDYrLjd70Qwaw true 4.5 118
WI Madison Vilas WUfQL5uURZwFootmgAMCyA true 4.0 226
WI Madison Vilas XgNy4vCHn-1NVzu6xgZIXg true 4.0 509
WI Madison Vilas XsDFvUcFxnOE0w9zKv0-iA true 3.0 76
WI Madison Vilas jEf18KTlek4zRHdsV6kIPQ true 4.0 349
WI Madison Vilas l_0JhWjplZBesMEIDAM4Fg true 4.5 527
WI Madison Vilas mJFNRws-E-ak1_CoBo9rcA false 4.5 22
WI Madison WBeltlineHwy 1vXBMP10It4jzUWBM-lJgA true 3.0 22
WI Madison WBeltlineHwy 2WxwgDMKm8XW6grbuMsDwg false 2.5 25
WI Madison WBeltlineHwy JfTE_LnG2D8jYKxiWavfzg false 3.0 34
WI Madison WBeltlineHwy f0gIqoBYj7GKOh6Icl8vAg true 4.0 123
WI Madison WBeltlineHwy gW0k1iIF1xuuFcLzI_HbBA true 3.5 184
WI Madison WBeltlineHwy nNTQtbPkTluPt5S9_RBFQ true 4.5 18
```

Testing

Unit Test case for REQ-4: Favorable type of business in city

TEST CASE	EXPECTED OUTPUT	ACTUAL OUTPUT	Result
Read json object with empty values	Ignore current json object and move to next line	Ignored the current json object	Pass
Passing Null Key and value pairs to Reduce function	Null pointer Exception should be thrown	Null pointer Exception thrown	Pass
Passing a JSON Object of type check-in info to mapper function mapper2	(32TNAtF2XVRYr1W_0deWMw,S11)	(32TNAtF2XVRYr1W_0deWMw,S11)	Pass
Passing integer values as key - value pair to reduce function	Cannot cast Integer to Text Exception	Cannot Cast Integer to Text Exception	Pass
Passing Text key and value pair to mapper class	Cannot Cast LongWritable to Text	Cannot cast LongWritable to Text	Pass

Functional Test cases

Use Case	Function being tested	Initial system state	Input	Expected Output	Actual Output
Existing Business Owners	Checking at what hours the business is busier	Home page is displayed	Choose Existing Business owners category and search for his business id. Click on this functionality	Output got displayed in the form of graphs	Output got displayed in the form of graphs
Future Business Owners	Favorable type of business in a city	Home page is displayed	Choose Future Business owner category. Search for the functionality and click on it	Output got displayed in the form of graphs	Output got displayed in the form of graphs
Government Organizations	Identify the cultural trends	Home page is displayed	Choose Government Organization and click on one of the cultural trends	Results will be displayed in any pictorial diagram format	Results will be displayed in any pictorial diagram format

Project Status

Current Status

- Out of 10 functional requirements, we have completed 9 requirements
- We are not yet created web application to present the output.

Future Enhancements

- We can use review text to extract more information and knowledge. NLP (Natural language processing) can be used to read review text and analyze it.
- Our system is mainly focused on future and existing business owners. But as professor suggested, more functionalities can be included for end users too.

Lessons Learned

New technology: We implemented project with hadoop map reduce algorithm which is relatively new technology. We learnt it starting from installation, sample tutorials and then implemented our project.

Contribution of CSC 230 course to this project

Proper documentation: Coursework helped us to prepare better documentation for our project which includes identifying functional – nonfunctional requirements, System diagram and detailed design (Use case, Sequence, Class diagram).

Automated test case: We learnt how to use Junit for creating automated test case which can be used over and over to test same functionalities without much effort. That helped us to write MR unit automated test case for our work.

References

- (1) <http://hadoop.apache.org/docs/current/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html>
- (2) http://www.yelp.com/dataset_challenge/
- (3) <http://www.ece.rutgers.edu/~marsic/books/SE/projects/ParkingLot/2012-g3-report3.pdf>