Requirements Analysis Document

NOVEMBER 21, 2022

BBB (Better Business Bureau)
Authored by: Data Divas (Medhanit Asrat Bekele, Mohammed Ahnaf Khalil, Wengel Tsegaselassie, Wen Sun)

TABLE OF CONTENTS

1. Introduction

The result of the requirement elicitation and the analysis activities are documented in the Requirements Analysis Document (RAD). The RAD describes the project in terms of data, known problems in the database as well as consequences associated with each type of problem.

1.1. Scope of project

In BBB, data management is normally done manually, which is a process that's not feasible for cost and scale reasons. With our help, BBB hopes to convert that lengthy process to a more authomatic and efficient process. Our project focuses on finding and implementing effective and powerful methods to improve data quality at BBB. Below is the summary of the scope of the project.

- Requirements, problem analysis and documentation
- Implementation of experiment/ test environment that accesses client data
- Writing an algorithm that solves specific data quality issues.
- Proof of solutions through well-documented testing procedures and results
- Document patterns and types of problems observed during the project

1.2. Objective

The project team will automate or semi automate the process of finding and correcting problems in BBB's database.

2. Current system

2.1. Blue Database

The current database system that BBB is operating on is called blue database. Blue database is a third-party database which is an on-disk, special-purpose datastore optimized for time-based data retrieval using small amounts of memory.

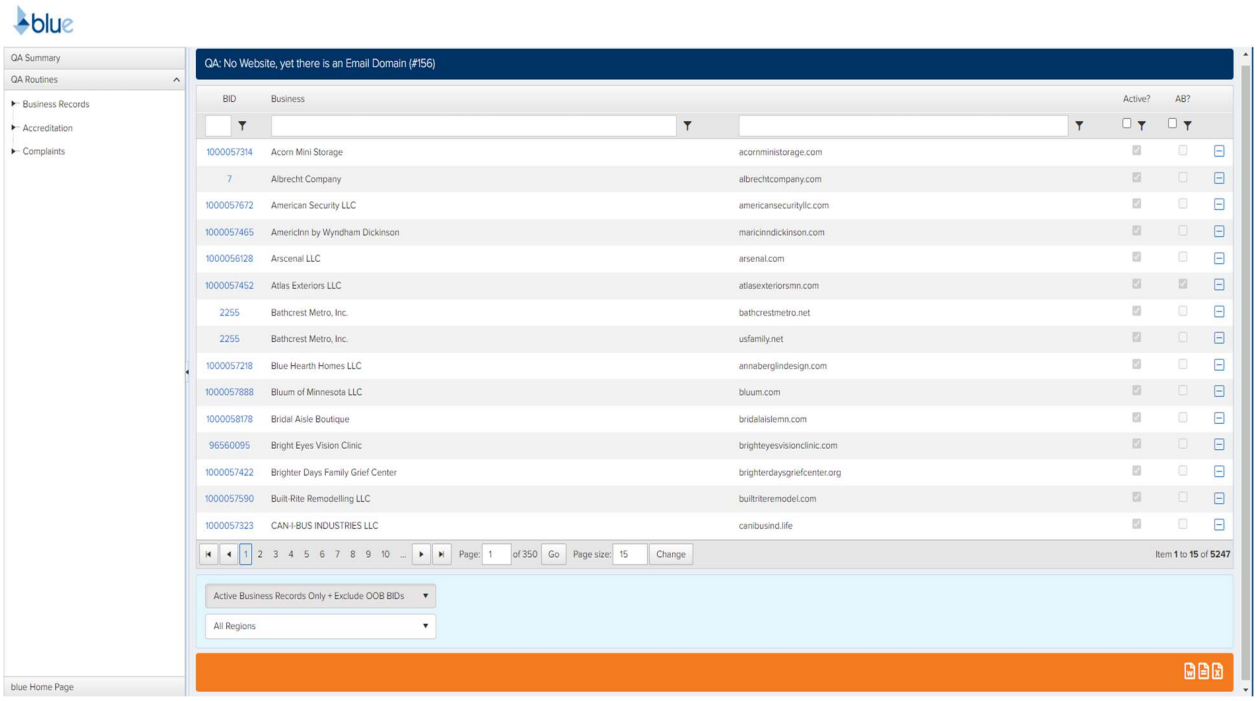


Figure 2: An example of what BBB's database (Blue database) looks like.

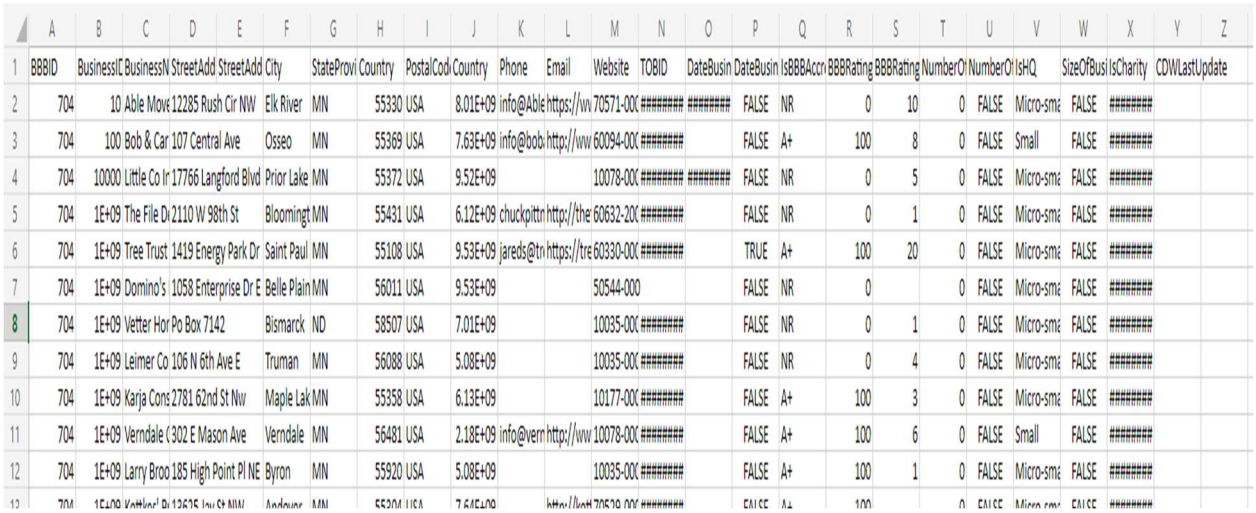
2.2. Exporting the data into a CSV file

Since we didn't want the modifications, we make to go live, our client decided to export the data from the database above into a csv file, which makes it easier to manipulate without making permanent changes in the database.

2.2.1. Type of data in the database

For our project we are mainly focusing on businesses located in Minnesota. Our current database consists of three tables each holding specific information about businesses in Minnesota. Each table will be explained in detail below.

1. Business Table



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1	BBBID	BusinessID	BusinessN	StreetAdd	StreetAdd	City	StateProvi	Country	PostalCod	Country	Phone	Email	Website	TOBID	DateBusin	DateBusin	IsBBBAccr	BBBRating	BBBRating	NumberOf	NumberOf	IsHQ	SizeOfBusi	IsCharity	CDWLastUpdate	
2	704	10	Able Move	12285	Rush Cir NW	Elk River	MN	55330	USA	8.01E+09	info@Able	https://ww	70571-00(	########	########	FALSE	NR	0	10	0	FALSE	Micro-sma	FALSE	########		
3	704	100	Bob & Car	107	Central Ave	Osseo	MN	55369	USA	7.63E+09	info@bob	http://ww	60094-00(	########		FALSE	A+	100	8	0	FALSE	Small	FALSE	########		
4	704	10000	Little Co Ir	17766	Langford Blvd	Prior Lake	MN	55372	USA	9.52E+09			10078-00(	########	########	FALSE	NR	0	5	0	FALSE	Micro-sma	FALSE	########		
5	704	1E+09	The File Di	2110	W 98th St	Bloomingt	MN	55431	USA	6.12E+09	chuckpitn	http://the	60632-20(	########		FALSE	NR	0	1	0	FALSE	Micro-sma	FALSE	########		
6	704	1E+09	Tree Trust	1419	Energy Park Dr	Saint Paul	MN	55108	USA	9.53E+09	jareds@tr	https://tre	60330-00(	########		TRUE	A+	100	20	0	FALSE	Micro-sma	FALSE	########		
7	704	1E+09	Domino's	1058	Enterprise Dr E	Belle Plain	MN	56011	USA	9.53E+09			50544-000			FALSE	NR	0		0	FALSE	Micro-sma	FALSE	########		
8	704	1E+09	Vetter Hor	Po Box 7142		Bismarck	ND	58507	USA	7.01E+09			10035-00(	########		FALSE	NR	0	1	0	FALSE	Micro-sma	FALSE	########		
9	704	1E+09	Leimer Co	106	N 6th Ave E	Truman	MN	56088	USA	5.08E+09			10035-00(	########		FALSE	NR	0	4	0	FALSE	Micro-sma	FALSE	########		
10	704	1E+09	Karja Cons	2781	62nd St Nw	Maple Lak	MN	55358	USA	6.13E+09			10177-00(	########		FALSE	A+	100	3	0	FALSE	Micro-sma	FALSE	########		
11	704	1E+09	Verndale (	302	E Mason Ave	Verndale	MN	56481	USA	2.18E+09	info@vern	http://ww	10078-00(	########		FALSE	A+	100	6	0	FALSE	Small	FALSE	########		
12	704	1E+09	Larry Broo	185	High Point Pl NE	Byron	MN	55920	USA	5.08E+09			10035-00(	########		FALSE	A+	100	1	0	FALSE	Micro-sma	FALSE	########		
13	704	1E+09	Kettler Pi	12625	Iw St NW	Andover	MN	55304	USA	7.64E+09		http://ww	70539-00(	########		FALSE	A+	100		0	FALSE	Micro-sma	FALSE	########		

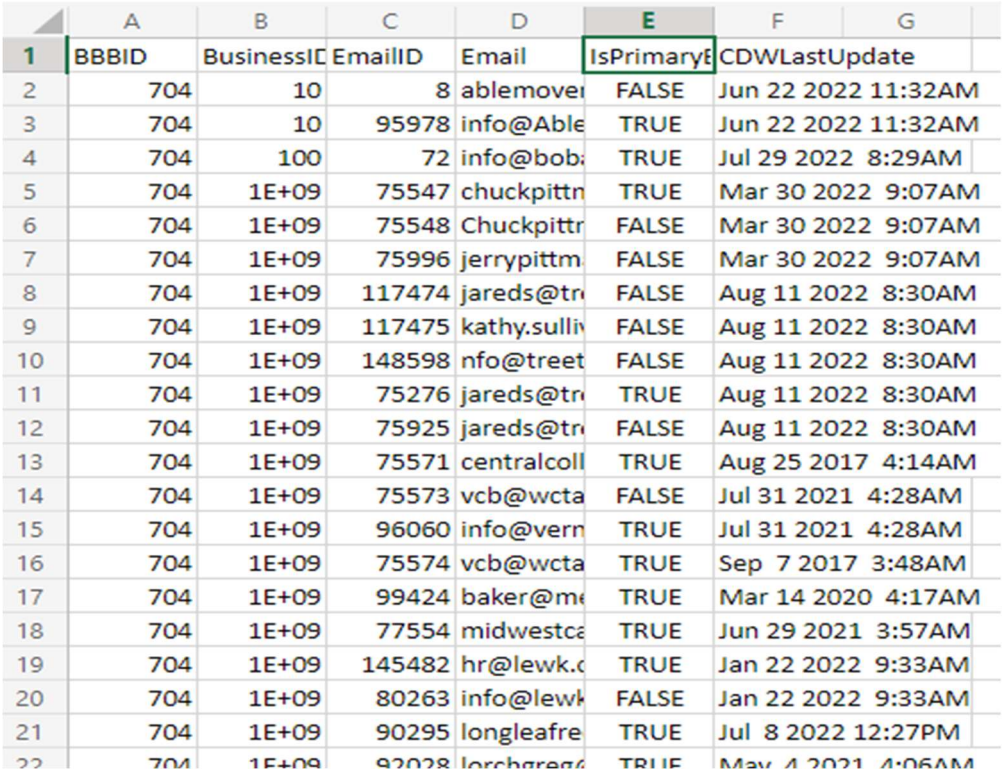
Figure 3: A table consisting of all the BBB accredited businesses in Minnesota along with all the necessary information. All the information included in that table are as follows:

Information in the table (Columns)	Type of data	Description
BBBID	Integer	Identifies businesses from Minnesota
BusinessID	Integer	The **primary key** for this table used to identify each business
BusinessName	String	Name of business
StreetAddress	String	The street address of the business
City	String	The city where the business is located
stateProvince	String	The state where the business is located
PostalCode	Integer	The postal code of the business
Country	String	The country where the business is located

Phone	Integer	Phone number associated with the business
Email	String	Email address associated with the business
Website	String	The website associated with the business
TOBID	String	
DateBusinessStarted	String	The date when the business started
DateBusinessClosed	String	The date when the business closed (if it is closed)
is BBB Accredited	String (true/false)	Whether the business is BBB accredited or not
BBBRatingGrade	String	Company's rating grade
BBBRatingScore	String	Company's rating score
NumberOfEmployees	Integer	Number of full-time employees in the business
NumberOfPartTimeEmployees	Integer	Number of part time employees in the business
IsHQ	String (true/false)	Is it the head quarter
SizeOfBusiness	Integer	The size of business
IsCharity	String (true/false)	Is the company for charity

CDWLastUpdate	String	The last time the database was updated

2. Emails Table



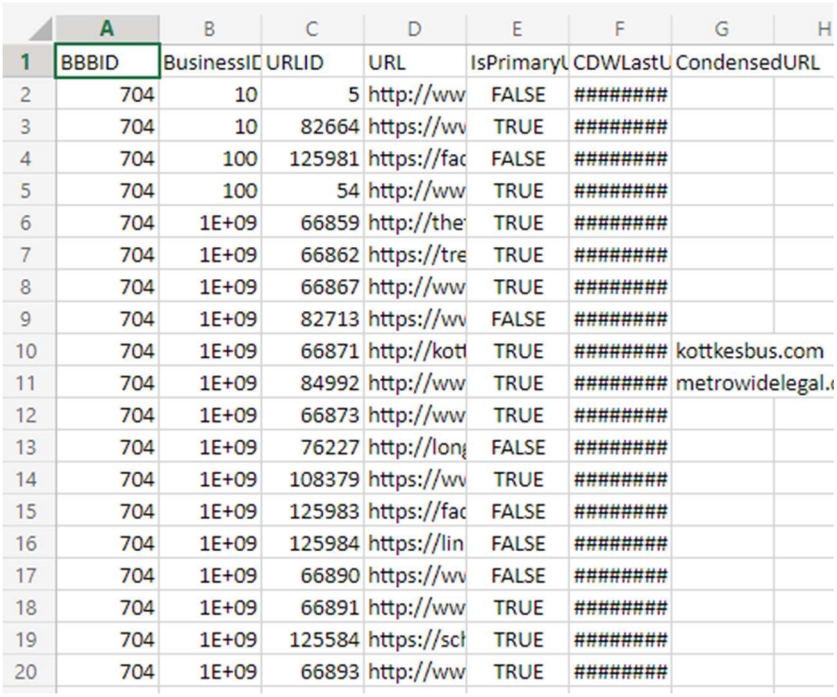
	A	B	C	D	E	F	G
1	BBBID	BusinessID	EmailID	Email	IsPrimaryE	CDWLastUpdate	
2	704	10	8	ablemover	FALSE	Jun 22 2022 11:32AM	
3	704	10	95978	info@Able	TRUE	Jun 22 2022 11:32AM	
4	704	100	72	info@bob:	TRUE	Jul 29 2022 8:29AM	
5	704	1E+09	75547	chuckpittn	TRUE	Mar 30 2022 9:07AM	
6	704	1E+09	75548	Chuckpittr	FALSE	Mar 30 2022 9:07AM	
7	704	1E+09	75996	jerrypittm	FALSE	Mar 30 2022 9:07AM	
8	704	1E+09	117474	jareds@tr	FALSE	Aug 11 2022 8:30AM	
9	704	1E+09	117475	kathy.sulli	FALSE	Aug 11 2022 8:30AM	
10	704	1E+09	148598	nfo@treet	FALSE	Aug 11 2022 8:30AM	
11	704	1E+09	75276	jareds@tr	TRUE	Aug 11 2022 8:30AM	
12	704	1E+09	75925	jareds@tr	FALSE	Aug 11 2022 8:30AM	
13	704	1E+09	75571	centralcoll	TRUE	Aug 25 2017 4:14AM	
14	704	1E+09	75573	vcb@wcta	FALSE	Jul 31 2021 4:28AM	
15	704	1E+09	96060	info@vern	TRUE	Jul 31 2021 4:28AM	
16	704	1E+09	75574	vcb@wcta	TRUE	Sep 7 2017 3:48AM	
17	704	1E+09	99424	baker@me	TRUE	Mar 14 2020 4:17AM	
18	704	1E+09	77554	midwestca	TRUE	Jun 29 2021 3:57AM	
19	704	1E+09	145482	hr@lewk.c	TRUE	Jan 22 2022 9:33AM	
20	704	1E+09	80263	info@lewk	FALSE	Jan 22 2022 9:33AM	
21	704	1E+09	90295	longleafre	TRUE	Jul 8 2022 12:27PM	
22	704	1E+09	92028	lorchgreg	TRUE	May 4 2021 4:06AM	

Figure 4: A table consisting of the emails associated with each business. All the information included are listed below:

Information in the table (Column)	Type of data	Description
BBBID	Integer	Identifies businesses from Minnesota
BusinessID	Integer	The **foreign key** for this table used to identify each business
EmailID	Integer	The **primary key** for this table used to identify the

		email of each business
Email	String	Email address associated with the businesses
IsPrimaryEmail	Boolean	Shows if the email is the primary email of the business
CDWLastUpdate	String	Lat time the database was updated

3. URL Table



	A	B	C	D	E	F	G	H
1	BBBID	BusinessID	URLID	URL	IsPrimaryU	CDWLastU	CondensedURL	
2	704	10	5	http://ww	FALSE	########		
3	704	10	82664	https://wv	TRUE	########		
4	704	100	125981	https://fac	FALSE	########		
5	704	100	54	http://ww	TRUE	########		
6	704	1E+09	66859	http://the	TRUE	########		
7	704	1E+09	66862	https://tre	TRUE	########		
8	704	1E+09	66867	http://ww	TRUE	########		
9	704	1E+09	82713	https://wv	FALSE	########		
10	704	1E+09	66871	http://kot	TRUE	########	kottkesbus.com	
11	704	1E+09	84992	http://ww	TRUE	########	metrowidelegal.(	
12	704	1E+09	66873	http://ww	TRUE	########		
13	704	1E+09	76227	http://lon	FALSE	########		
14	704	1E+09	108379	https://wv	TRUE	########		
15	704	1E+09	125983	https://fac	FALSE	########		
16	704	1E+09	125984	https://lin	FALSE	########		
17	704	1E+09	66890	https://wv	FALSE	########		
18	704	1E+09	66891	http://ww	TRUE	########		
19	704	1E+09	125584	https://sch	TRUE	########		
20	704	1E+09	66893	http://ww	TRUE	########		

Figure 5: A table consisting of all the URLs of the BBB accredited businesses in Minnesota along with all the necessary information. All the information included in that table are as follows:

Information in the table (Column)	Type of data	Description
BBBID	Integer	Identifies businesses from Minnesota

BusinessID	Integer	The **foreign key** of this table that Identifies each business from Minnesota
URLID	Integer	The **primary key** of this table that identifies the URL of the business
URL	String	Identifies the URL of the business
IsPrimaryURL	String (true/false)	Identifies whether the URL is the primary URL
CDWLastUpdate	String	The last time the database was updated
CondensedURL	String	

3. Analysis of the known problems in the data and their consequences

Our exported CSV file contains more than 150,000 businesses along with their necessary information. Since that file is too large, we decided to take a tranche (Sample) of the data to test our code. To get the tranche of the data, we used a python package called pandas. Our tranche consisted of 100 rows from the table. We chose 100 rows because it was the ideal size to cover the majority of test cases without impeding execution.

Common problems observed in the URL until now:

- Syntax error
- Redirect of business website URL
- Page not found
- Site cannot be reached

3.1. Evaluated each URL for possible syntax problems

We evaluated the syntax for each URL from our tranche by using regex. We checked if the given URLs matched the expected pattern for a common URL. So far, we have observed unnecessary special characters at the end of each URL. We expect to see more as we increase our data set.

3.2. Identified the status code associated with each URL

The second step in the analysis of data was to write a code that reads each URL from our tranche and evaluates the status codes that are associated

with each URL. We recorded each URL with their observed errors in a different CSV file for easier testing and observation.

	A	B	C	D
1	bid	url_id	url	status_code
2	100	125981	https://facebook.com/pages/category/automotive--aircraf	200
3	100	54	http://www.bobandcarls.com	200
4	1000000001	66862	https://treetrust.org/	200
5	1000000007	66867	http://www.verndalecustombuilders.com/	200
6	1000000007	82713	https://www.facebook.com/Verndale-Custom-Builders-11	200
7	1000000010	66871	http://kottkesbus.com/	200
8	1000000011	84992	http://www.metrowidelegal.com/	200
9	1000000012	66873	http://www.midwestcanine.com/	200
10	1000000019	108379	https://www.discountplumbers.com	200
11	1000000019	125983	https://facebook.com/discount-plumbing-and-drain-cleani	200
12	1000000019	66890	https://www.discountrooter.com	200
13	1000000023	125584	https://schaperengineering.com/	200
14	1000000026	66893	http://www.insurancespecialiststeam.com	200
15	1000000028	66897	http://www.justicelawncare.com/	200
16	1000000035	66901	https://www.facebook.com/kleinlawnandlandscape/	200
17	1000000036	82204	https://www.facebook.com/angocc/	200
18	1000000040	82726	http://www.robbinsdalemn.com/services/municipal-liquo	200
19	1000000042	66912	http://www.heritagexteriors.com	200
20	1000000043	73116	http://www.diyturfguru.com	200
21	1000000049	66919	http://www.belovedlasertattooremoval.com	200
22	1000000051	66925	http://www.stampnstorage.com/	200
23	1000000052	111896	http://personalizedinsuranceconsulting.com/	200

Figure 6: URLs with status code 200 “OK” response

When a URL returns a status code 200 it means that the request has succeeded. The client has requested documents from the server. The server has replied to the client and given the client the documents. This is the status code that is the most favorable.
We haven’t found any data with other status codes in the range 201-299 till now.

11	1000000050	66920	http://www.np-cpa.com	301
12	1000000051	66925	http://www.stampnstorage.com/	301
13	1000000052	111896	http://personalizedinsuranceconsulting.com/	301
14	1000000056	66928	http://www.traversecc.org/	301
15	1000000062	86651	https://www.facebook.com/Ace-Insurance-Age	301
16	1000000064	66931	http://www.hairstudio763.com	302
17	1000000065	66933	http://www.furnituredealer.net	301
18	1000000083	66952	http://www.firstpitch.com	301
19	1000000084	74240	http://www.rwplumb.com/	301
20	1000000095	66958	http://www.photographermn.net	301
21	1000000101	67022	http://www.plumbercontractorsduluth.com	301
22	1000000103	66964	http://martinizing.com/xerxes/default.aspx	301
23	1000000105	66974	http://www.fswchiro.com	301
24	1000000107	66967	http://primeattach.com/	301
25	1000000120	66973	http://www.cielloftandhome.com/	301
26	1000000123	66978	http://doggonecutepuppies.com/	302
27	1000000127	66983	https://www.facebook.com/Marks-Aerial-Serv	301

Figure 7: URLs with status Code 300 “Redirect” response

We received status code 301 and 302 from the tranche we tested. Other status codes might be received if the testing pool was larger. 300 status codes are usually for URL redirects. Redirects are the result of change, and this can be a change of SEO (Search Engine Optimization) strategy, business changes, change in policy and many more.

- **301**- The requested resource has been permanently moved to the URL given by the location headers.
- **302**- The resource requested has been temporarily moved to the URL given by the location header.

Errors associated with 300 status codes could cause multiple problems as:

- Windows might run slow and respond slowly
- Freezing of system for few seconds

Redirects are not a high priority issue because a 300 response is cacheable by default. Google normally handles this error. For example, when Google notices 301 redirects after indexing a site, its spiders update the site index and restore the page ranking.

	A	B	C	D
1	bid	url_id	url	status_code
2	1000000000	66859	http://thefiledepot.com	406
3	1000000020	66891	http://www.ronsautorepairmanka	403
4	1000000033	66899	http://www.abcbookwerks.com	406
5	1000000046	66914	http://www.caliberproductsinc.co	403
6	1000000048	66915	http://www.shackslaw.com/	406
7	1000000050	66920	http://www.np-cpa.com	403
8	1000000076	66941	http://creativeapestudios.com	406
9	1000000103	66964	http://martinizing.com/xerxes/de	404
10	1000000115	66970	http://www.linkconsultinggroup.c	404
11	1000000119	124112	https://www.igtrestore.com/	403
12	1000000136	66990	http://www.ccsconcrete.com/inde	404
13	1000000144	67009	http://festlerlandsurveying.com/	403
14	1000000147	67001	http://aspengarage.com	406
15	1000000147	73957	http://www.AspenGarage.com	406
16	1000000157	68886	http://www.mnintegrity.com	404
17	1000000160	67007	http://www.eyeofthetigerdefense	404
18	1000000180	67052	http://theramg.com/	403
19	1000000185	67035	http://www.freshthyme.com	403
20	1000000188	74033	https://www.hsdoorsystems.com/	403

Figure 8: URLs with status code 400 “bad request” response

We received status code 403, 404 and 406 status code from the tranche we tested. Other status codes might be received if the testing pool was larger. 400 status codes are usually for bad response. It indicates that the server cannot or will not process the request due to a client error.

- **403**- The server understands the request but refuses to authorize it as the request is a third party one. Websites with status code 403 will not pose any issue for the user or

BBB. It was just received by us as the websites refuse scanning of their websites due to third-party domains.

- **404**- Page not found. These websites are the ones that will pose a huge problem for users and BBB. Either the business is closed or the information in BBB database is incorrect.
- **406**- Not acceptable (Details not yet available)

	A	B	C	D
1	bid	url_id	url	status_code
2	1000000019	125984	https://linkedin.com/in/discountplumbers	999

Figure 9: URL with status code greater than 499:

Received only 1 such record from the 100 records, however there might be more as we go ahead with the project. The URL with this status code doesn't specifically pose any problem for the company or the user so the error we received is for websites such as "LinkedIn" do not allow scanning of their websites.

- **999**- It is used as a "catch-all" error code presented when a more specific error code is not provided by the server we are trying to access.

	A	B	C	D
1	bid	url_id	url	status_code
2	10	82664	https://www.ablemovingcorp.com/	-1
3	1000000015	76227	http://longleafreclaimedtimber.com/	-1
4	1000000032	75725	http://www.acleanhomemn.com	-1
5	1000000055	66927	http://www.forpetssake.rocks	-1
6	1000000086	66953	https://www.americanequity.com/minneapolis-minn	-1
7	1000000099	74197	http://www.alpineaircare.com	-1
8	1000000131	67021	http://www.executivelandscapingmn.com/	-1
9	1000000139	118009	https://www.badlandsdoitbesthardware.com	-1
10	1000000156	67005	http://www.knudsenlandscapingandexcavating.com/	-1
11	1000000164	67053	http://www.precisionhomeconstruction.com/	-1
12	1000000171	67057	http://parlorseven.com/home/	-1
13				

Figure 10: URL's which can't be reached

In our code we returned "-1" status code if the site cannot be reached. The problem that the company or users face is massive as either the company is closed, or the information BBB has in their database for that specific company is faulty.

4. Approvals

Approver/Reviewer	Role	Date of Approval/ Reviewal
Ryan Sharp	Main Client	
Eli Johnson	Technical Specialist	
Lin Chase	Faculty Coach	