Sagarkumar .P. Vaghasia
B00878629

CSCI 5408 – Data Management, Warehousing and Analytics
Assignment 1
Problem 1

I visited https://resources-covid19canada.hub.arcgis.com/search for identifying key entities, attributes and creating an ERD.

- It collects daily data on COVID-19 cases, deaths, recoveries, testing and vaccinations at the health region and province levels.
- There are in total 131 datasets on this link https://resources-covid19canada.hub.arcgis.com/search?collection=Dataset , but I will be focusing on datasets of COVID-19 case details by province.
- Data is collected from publicly available sources such as government datasets and news releases. Here, I am taking datasets of B.C.COVID-19 Collection Centres.

Entities	**Source URL**	**Attributes**
B.C._Health_Authority_Boundaries__with_Provincial_Health_Services_Boundary_	https://resources-covid19canada.hub.arcgis.com/datasets/bcgov03::b-c-health-authority-boundaries-with-provincial-health-services-boundary/about	HA_ID HA_Name HA_Pop20 Date_Updat Source URL GlobalID Shape_Area Shape__Length
B.C._COVID-19_-_Case_Details	https://resources-covid19canada.hub.arcgis.com/datasets/bcgov03::b-c-covid-19-case-details/about	Reported_Date HA Sex Age_Group

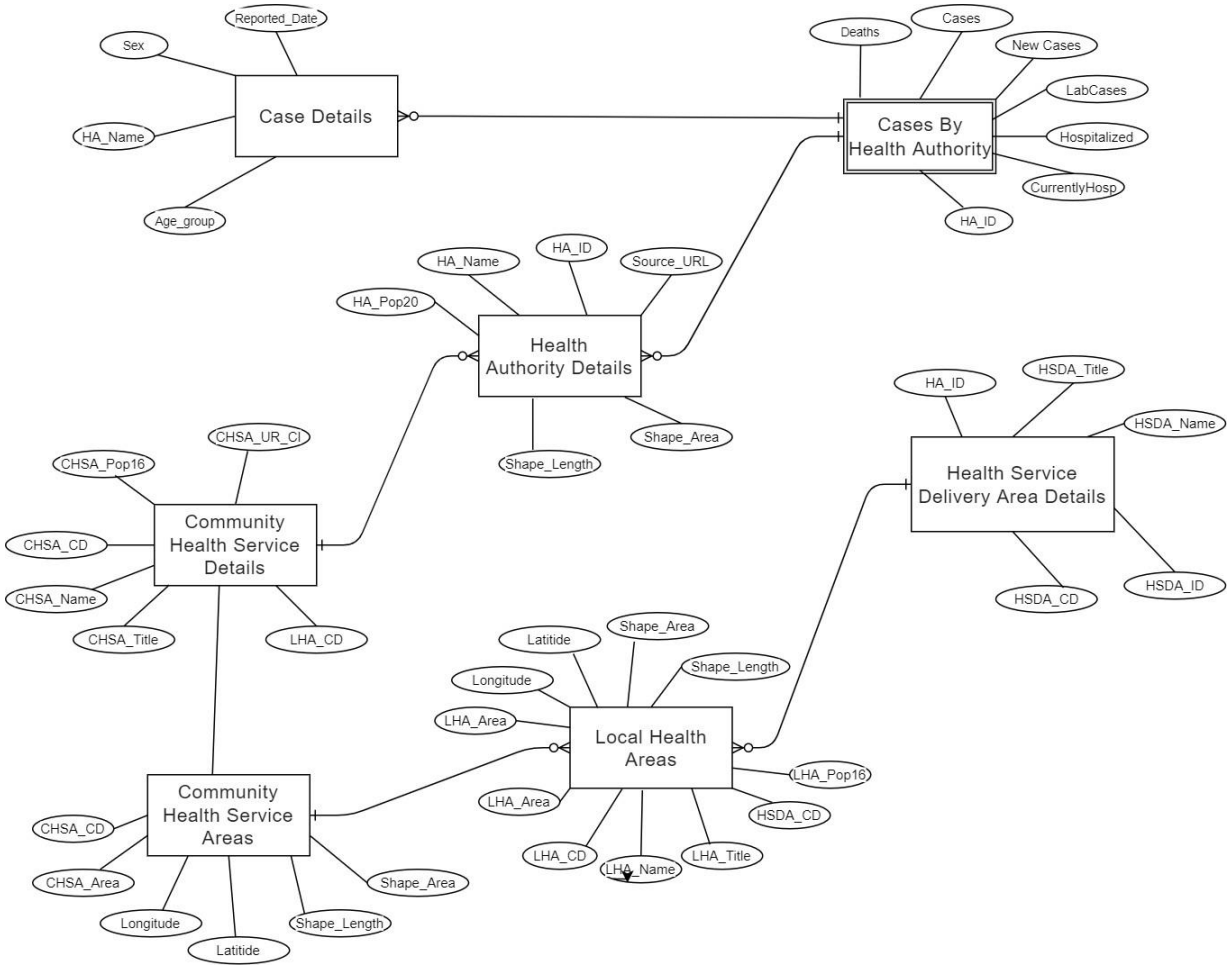
B.C._COVID-19_Cases_by_Health_Authority	https://resources-covid19canada.hub.arcgis.com/datasets/bcgov03::b-c-covid-19-cases-by-health-authority/about	HA_ID Health Authority Health Authority Population Deaths COVID-19 Cases Update Date Source Shape_Area Shape_Length Total Hospitalized to Date Currently Hospitalized Currently Admitted to ICU New Cases Laboratory Cases EpiCases New Deaths
BC_Local_Health_Areas	https://resources-covid19canada.hub.arcgis.com/datasets/exchange::bc-local-health-areas/about	LHA_CD LHA_Name LHA_Title LHA_CD1997 HSDA_CD HSDA_ID HSDA_Name HSDA_Title HA_CD HA_ID HA_Name HA_Title

		LHA_Pop16 LHA_Area Latitude Longitude Shape_Area Shape_Length Latitude Longitude Shape_Area Shape_Length HEALTH_SERVICE_DLVR_AREA_NAME LOCAL_HEALTH_AREA_CODE LOCAL_HEALTH_AREA_NAME Hip/Knee Surgery CreationDate Creator EditDate Editor
BC_Community_Health_Service_Areas	https://resources-covid19canada.hub.arcgis.com/datasets/exchange::bc-community-health-service-areas/about	CHSA_CD CHSA_Name CHSA_Title LHA_CD LHA_Name LHA_Title LHA_CD1997 HSDA_CD HSDA_ID HSDA_Name

		HSDA_Title HA_CD HA_ID HA_Name HA_Title CHSA_UR_Cl CHSA_Pop16 CHSA_Area Latitude Longitude Shape_Area Shape_Length HEALTH_SERVICE_DLVR_AREA_NAME LOCAL_HEALTH_AREA_CODE LOCAL_HEALTH_AREA_NAME Hip/Knee Surgery CreationDate Creator EditDate Editor

Chen-model:

After identifying entities and defining attributes, I created initial Chen model for the conceptual which I have attached below:



Normalization

Considering unnormalized datasets of BC_Community_Health_Service_Areas.csv

BC_Community_Health_Service_Areas

CHSA_CD	CHSA_Name	CHSA_Title	LHA_CD	LHA_Name	LHA_Title	LHA_CD1997	HSDA_CD	HSDA_ID	HSDA_Name	HSDA_Title	HA_CD	HA_ID	HA_Name	HA_Title	CHSA_UR_CI	CHSA_Pop16	CHSA_Area	Latitude
1110	Fernie	1110 Fernie	111	Fernie	111 Fernie	1	11	11 EK	East Kootenay	11 East Kootenay	1	1 IHA	Interior	1 Interior	5 Rural Hub	15531	8043.8	49.417397
1120	Cranbrook	1120 Cranbrook	112	Cranbrook	112 Cranbrook	2	11	11 EK	East Kootenay	11 East Kootenay	1	1 IHA	Interior	1 Interior	4 Small Urban	26248	4473.71	49.552631
1130	Kimberley	1130 Kimberley	113	Kimberley	113 Kimberley	3	11	11 EK	East Kootenay	11 East Kootenay	1	1 IHA	Interior	1 Interior	6 Rural	9178	4345.66	49.804307
1140	Windermere	1140 Windermere	114	Windermere	114 Windermere	4	11	11 EK	East Kootenay	11 East Kootenay	1	1 IHA	Interior	1 Interior	6 Rural	9487	10980.52	50.534955
1150	Creston	1150 Creston	115	Creston	115 Creston	5	11	11 EK	East Kootenay	11 East Kootenay	1	1 IHA	Interior	1 Interior	6 Rural	12634	3793.89	49.247658
1160	Golden	1160 Golden	116	Golden	116 Golden	18	11	11 EK	East Kootenay	11 East Kootenay	1	1 IHA	Interior	1 Interior	6 Rural	6851	13347.11	51.60957
1210	Kootenay Lake	1210 Kootenay Lake	121	Kootenay Lake	121 Kootenay Lake	6	12	12 KB	Kootenay Boundary	12 Kootenay Bounda	1	1 IHA	Interior	1 Interior	6 Rural	3209	6543.23	50.29662
1220	Nelson	1220 Nelson	122	Nelson	122 Nelson	7	12	12 KB	Kootenay Boundary	12 Kootenay Bounda	1	1 IHA	Interior	1 Interior	6 Rural	25523	4809.42	49.502751
1230	Castlegar	1230 Castlegar	123	Castlegar	123 Castlegar	9	12	12 KB	Kootenay Boundary	12 Kootenay Bounda	1	1 IHA	Interior	1 Interior	5 Rural Hub	13710	1942.58	49.398678
1240	Arrow Lakes	1240 Arrow Lakes	124	Arrow Lakes	124 Arrow Lakes	10	12	12 KB	Kootenay Boundary	12 Kootenay Bounda	1	1 IHA	Interior	1 Interior	6 Rural	4501	7391.98	50.241121
1250	Trail	1250 Trail	125	Trail	125 Trail	11	12	12 KB	Kootenay Boundary	12 Kootenay Bounda	1	1 IHA	Interior	1 Interior	4 Small Urban	19367	1131.23	49.124559
1260	Grand Forks	1260 Grand Forks	126	Grand Forks	126 Grand Forks	12	12	12 KB	Kootenay Boundary	12 Kootenay Bounda	1	1 IHA	Interior	1 Interior	6 Rural	8611	2691.95	49.34925
1270	Kettle Valley	1270 Kettle Valley	127	Kettle Valley	127 Kettle Valley	13	12	12 KB	Kootenay Boundary	12 Kootenay Bounda	1	1 IHA	Interior	1 Interior	6 Rural	3469	4347.28	49.413808
1310	Southern Okanagan	1310 Southern Okar	131	Southern Okanagan	131 Southern Okana	14	13	13 OK	Okanagan	13 Okanagan	1	1 IHA	Interior	1 Interior	5 Rural Hub	19231	1304.34	49.20982
1320	Penticton	1320 Penticton	132	Penticton	132 Penticton	15	13	13 OK	Okanagan	13 Okanagan	1	1 IHA	Interior	1 Interior	3 Medium Urban	41799	1562.86	49.535438
1330	Keremeos	1330 Keremeos	133	Keremeos	133 Keremeos	16	13	13 OK	Okanagan	13 Okanagan	1	1 IHA	Interior	1 Interior	6 Rural	5101	2479.93	49.222987
1340	Princeton	1340 Princeton	134	Princeton	134 Princeton	17	13	13 OK	Okanagan	13 Okanagan	1	1 IHA	Interior	1 Interior	5 Rural Hub	4781	4826.04	49.465568
1350	Armstrong/Spallumc	1350 Armstrong/Spa	135	Armstrong/Spallumc	135 Armstrong/Spall	21	13	13 OK	Okanagan	13 Okanagan	1	1 IHA	Interior	1 Interior	5 Rural Hub	10220	263.75	50.445902

The first step in normalizing a relation is to remove the repeating groups. In the above image all the details related to LHA_CD (Local health area) are repeated information of BC_Local_Health_Areas datasets. So, we can remove all the column related to LHA_CD Reference from the table BC_Community_Health_Service_Areas and link LHA_CD as reference key.

BC_Community_Health_Service_Areas

CHSA_CD	CHSA_Name	CHSA_Title	LHA_CD	CHSA_UR_CI	CHSA_Pop16	CHSA_Area	Latitude	Longitude	Shape__Area	Shape__Length
1110	Fernie	1110 Fernie	111	5 Rural Hub	15531	8043.8	49.417397	-114.917924	19065095661	1057610.
1120	Cranbrook	1120 Cranbrook	112	4 Small Urban	26248	4473.71	49.552631	-115.593391	10602419936	752612.41
1130	Kimberley	1130 Kimberley	113	6 Rural	9178	4345.66	49.804307	-116.055836	10421294217	675573.4
1140	Windermere	1140 Windermere	114	6 Rural	9487	10980.52	50.534955	-115.965026	2714819832 3	1131209.4
1150	Creston	1150 Creston	115	6 Rural	12634	3793.89	49.247658	-116.571432	8894436701	500296.56

The next step in normalization is No non-prime attribute is dependent on the proper subset of any candidate key of table.

However, in above table CSHA_TITLE,CHSA_Name,CSHA_UR_CI,CHSA_Pop16,CSHA_Area is dependent on CSHA_CD alone which is a proper subset of candidate key. To make the table complies we can disintegrate it in two tables like this:

Table-1

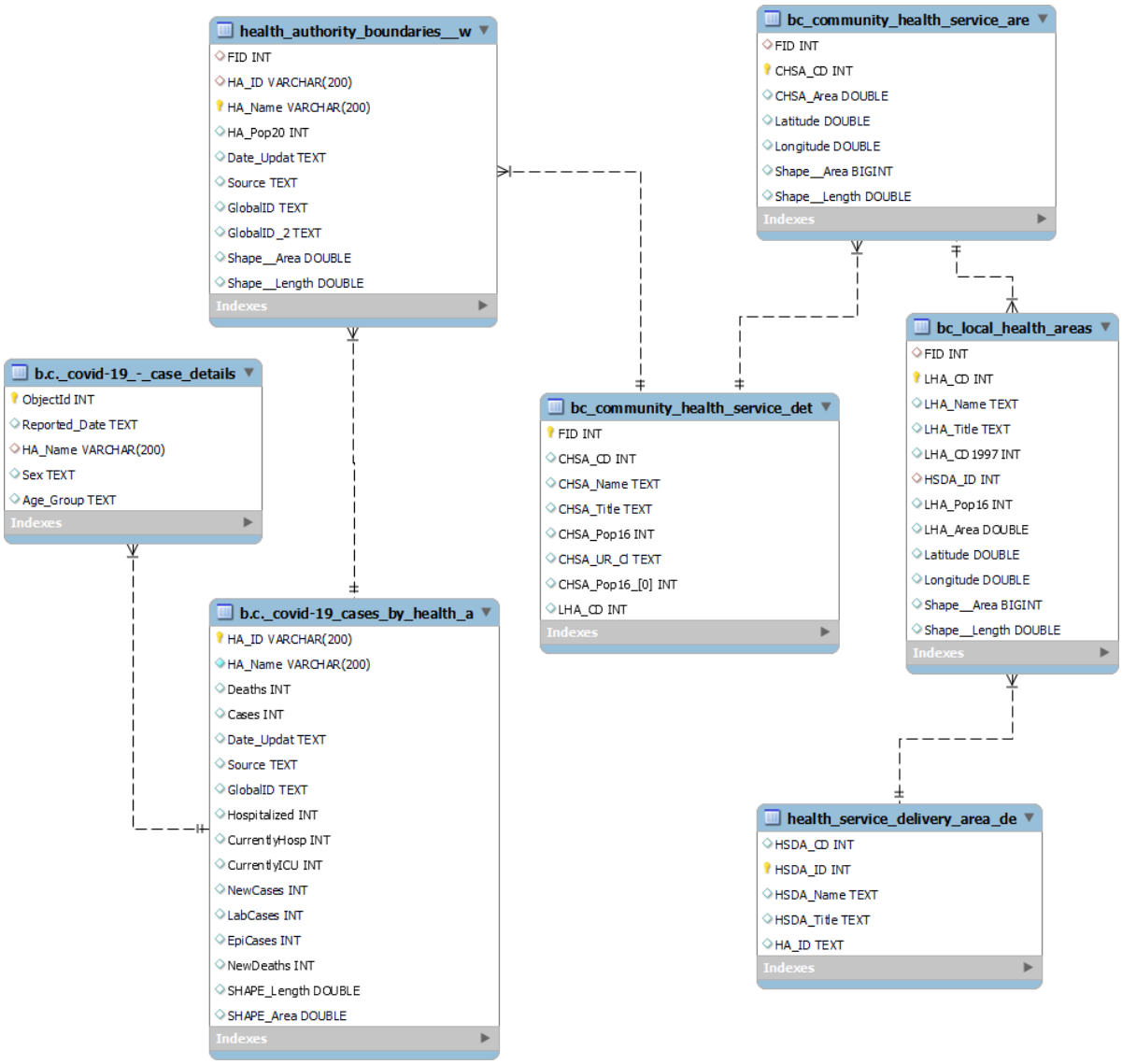
BC_Community_Health_Service_DETAILS

CHSA_CD	CHSA_Name	CHSA_Title	CHSA_UR_CI	CHSA_Pop16	LHA_CD
1110	Fernie	1110 Fernie	5 Rural Hub	15531	111
1120	Cranbrook	1120 Cranbrook	4 Small Urban	26248	112
1130	Kimberley	1130 Kimberley	6 Rural	9178	113
1140	Windermere	1140 Windermere	6 Rural	9487	114
1150	Creston	1150 Creston	6 Rural	12634	115

Table-2

BC_Community_Health_Service_Areas					
CHSA_CD	CHSA_Area	Latitude	Longitude	Shape__Area	Shape__Length
1110	8043.8	49.417397	-114.917924	19065095661	1057610.38
1120	4473.71	49.552631	-115.593391	10602419936	752612.4112
1130	4345.66	49.804307	-116.055836	10421294217	675573.421
1140	10980.52	50.534955	-115.965026	27148198323	1131209.462

After normalizing all the data, I created 7 tables where the data is stored. Then, I imported CSV files to the MySQL database where I assigned primary and foreign keys with the related tables and lastly by applying reverse engineering, I generated ERD in MySQL. I have attached the ERD below:



References :

[1] draw.io [online] : https://app.diagrams.net/

[2] Chen-model notations [online] : https://vertabelo.com/blog/chen-erd-notation/

“I Sagarkumar .P. Vaghasia, declare that in assignment 1 of CSCI 5408 course, data scrapping is not done programmatically or using any online or offline tools. However, the webpages or the domain mentioned in this document are visited manually, and some useful information is gathered for education purpose only. Information, such as email, personal contact numbers, or names of people are not extracted. The course instructor or the Faculty of Computer Science cannot be held responsible for any misuse of the extracted data”.