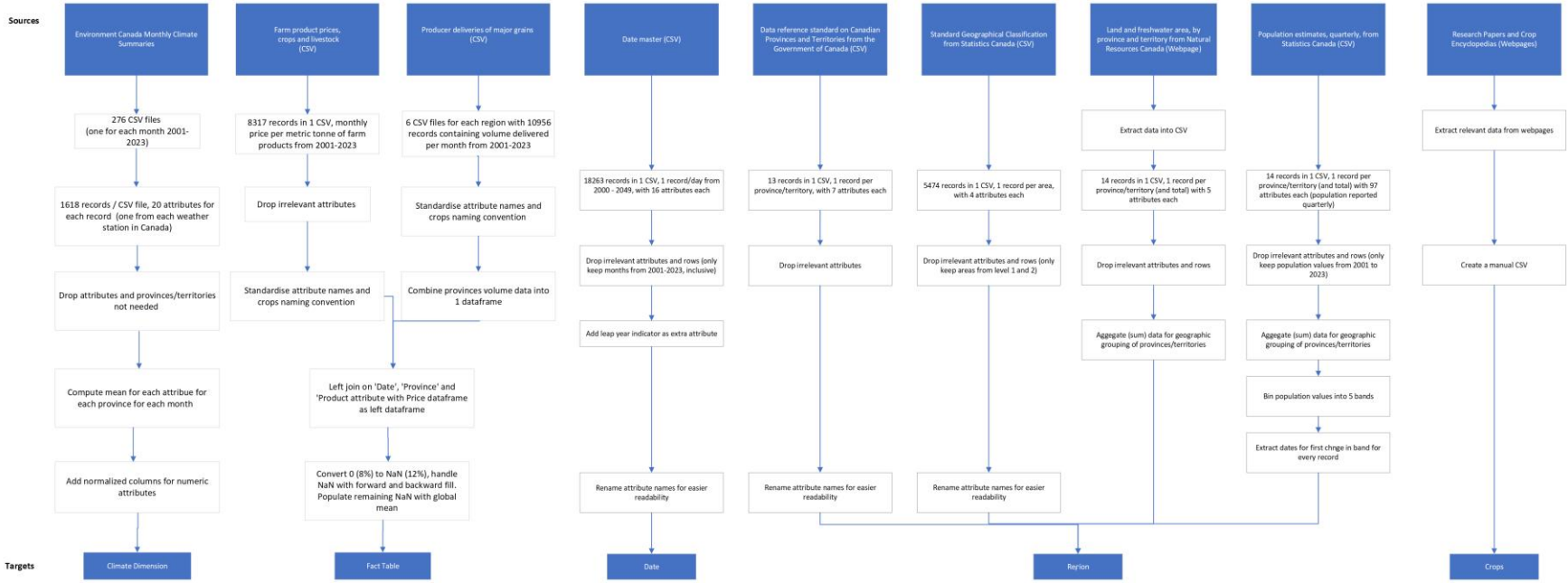
A. High-Level Data Staging Plan



B. Details

Climate Dimension:

The source of data is the Monthly Climate Summaries from Environment Canada, in the form of 276 CSV files, one for each month between January 2001 to December 2023. For each month, there is a record from each of the 1618 weather stations across Canada reporting 20 different weather attributes. The 5 columns ‘'DwTm', 'DwTx', 'DwTn', 'DwS', 'DwP', and 'DwBS' in the dataset were dropped as they are columns indicating the number of days temperature, snowfall, precipitation, and bright sunshine data was not available for a particular station, however, the columns are not meteorological attributes themselves. Datapoints from the provinces and territories of Northwest Territories, Nunavut, Yukon, Newfoundland and Labrador, Nova Scotia, New Brunswick, and Prince Edward Island were also dropped as there is insufficient crop data for those regions, so the regions are not included in the analysis.

The attributes from the original datasets were renamed for clarity:

Old Attribute Name	New Attribute Name
‘Tm'	'Mean_Temp'
'Tn'	'Min_Temp'
'Tx'	'Max_Temp'
'D'	'Mean_Diff_from_Normal'
‘S’	'Snowfall'
‘S%N’	'Percent_of_Normal_Snowfall'
‘P’	'Total_Precip'
'P%N'	'Percent_of_Normal_Precip'
‘S_G’	'Snow_on_Ground'
Pd	'Days_with_Precip'
BS	'Bright_Sunshine_Days'
BS%	'Percent_of_Normal_Bright_Sunshine'
'HDD'	'Heating_Degree_Days'
‘CDD’	'Cooling_Degree_Days'

The mean of each climate attribute across all the weather stations in a particular province was then computed for each province for each month. The values were then normalized to be in the range [0-1] with normalized values added as separate columns in the Climate Dimension to be available for future analysis

Date Dimension:

Data for the Date dimension was generated from a Date dimension generator, in the form of a singular CSV file with records for every day from Jan 1st, 2000 to Dec 31st, 2049. Each record contained the following information about the day: what month it belonged to (the month's properties), the quarter it belonged to (the quarter's properties) and the year it belonged to (the year's properties), totaling to 16 attributes. The columns 'CalendarMonthStartDate', ' CalendarMonthEndDate', 'CalendarQuarterStartDate', 'CalendarQuarterEndDate', 'CalendarYearEndDate', 'CalendarYearStartDate' were all dropped as they are not relevant to an analysis for monthly reporting. Additionally, the data cropped to only address the days from 2001 to 2023, and further truncated to only include the first day of every month – since the grain for the fact table is in months. The 'LeapYear' attribute was also added to every month and populated on values based on the 'CalendaNumberOfDaysInYear' to be used in further analyses. Lastly, the attributes were renamed for easier interpretation of their meaning and easier inferencing:

Old attribute name	New attribute name
'Id'	'Date_ID'
'MonthLongName'	'Month_LongName'
'MonthShortName'	'Month_ShortName'
'CalendarMonth'	'Month'
'CalendarNumberOfDaysInMonth'	'NumberOfDaysInMonth'
'CalendarQuarter'	'Quarter'
'CalendarNumberOfDaysInQuarter'	'NumberOfDaysInQuarter'
'CalendarYear'	'Year'
'CalendarNumberOfDaysInYear'	'NumberOfDaysInYear'

Region Dimension:

Data for the region dimensions was taken from several sources. The detailed breakdown is given in the table below:

Reference ID	Data	Source
1	Names, numeric codes, etc.	Government of Canada
2	Geographical classification	Statistics Canada
3	Land and freshwater area	Natural Resources Canada
4	Population estimates	Statistics Canada

Although data for [1], [2], and [4] were available in a downloadable CSV format, the data for [3] was only available on the webpage, requiring it to be manually extracted to a CSV format.

Starting with [1], there were a total of 13 records, one for each province/territory in Canada, with attributes like name in English and French, abbreviations in English and French, codes, etc. The 'iso_code' attribute was dropped as it would not provide any insight into the data and the rest of the attributes were renamed for better readability:

Old Attribute Name	New Attribute Name
'nm_en'	'Name_EN'
'nm_fr'	'Name_FR'
'ab_en'	'Abbrev_EN'
'ab_fr'	'Abbrev_FR'
'al_code'	'AlphaCode'
'nu_code'	'NumericCode'

Data provided from [2] contained information on the entire hierarchical structure of Canada, including geographical regions, provinces and territories, census divisions and census subdivisions. This information was stored in 5475 records with attributes like 'Level', 'Hierarchical structure', 'Code', and 'Class title' for each. However, as we are only interested in the geographical regions and provinces and territories for this analysis, the rest of the data had to be dropped, resulting in 20 remaining records. Using the first digit of the 'Code' attribute, the geographical region for every province and territory could be determined, and added to the region dimension under the 'GeorgraphicRegion' attribute. Additionally, the geographical regions from [2] were added as separate rows to the region dimension, which resulted in 3 new rows for 'Atlantic', 'Prairies', and 'Territories'.

The [3] data included information on the provinces and territories alongside their total area, land coverage and freshwater coverage. Therefore, in total, there were 14 records (including 'Canada') with their corresponding values in the CSV. The area percentage,

area in km, land in km, and freshwater in km were all added to the region dimension as new attributes. Data for the additional geographic regions from [2] was summed up and added to region dimension. 2 additional columns were created and calculated for all rows – 'LandOfArea_Percent' and 'FreshwaterOfArea_Percent' – to deepen the insight that could be gained from the region dimension.

Data from [4] included the population counts for every province and territory, including Canada, reported quarterly from Q1 2000 to Q4 2023. Data from before Q1 2001 was dropped, population values aggregated for the Atlantic, Prairies, and Territories regions, and population counts were binned into the following 5 categories: (0, 1000000], (1000000, 5000000], (5000000, 10000000], (10000000, 15000000], (15000000, 50000000]. Then only the dates for the changes in categories were kept, alongside the initial population count in Q1 2001, and the dates with their corresponding population band were merged into the region dimension.

Crop Dimension:

The data of the crop dimension is derived from many different research papers and encyclopedia on Canadian crops. Relevant information was gathered from the sources and used to formulate the crop dimensions. We chose this approach because our crop dimension consists mainly of static data which are not affected by time, therefore there are no datasets tracking the information we need for our crop dimension. The crop dimension consists of the crop's Common Name, Species, Family, Minimum Temperature (°C) and Maximum Temperature (°C).

Fact Table:

The data from the fact table is derived from the "Producer deliveries of major grains" and "Farm product prices, crops and livestock" datasets from Statistics Canada. 6 datasets were generated from "Producer deliveries of major grains" dataset for each province. There are 15 attributes, "REF_DATE", "GEO", "DGUID", "Type of grain", "UOM", 'UOM_ID", "SCALAR_FACTOR", "SCALAR_ID", "VECTOR", "COORDINATE", "VALUE", "STATUS", "SYMBOL", "TERMINATED", "DECIMALS". "REF_DATE", "GEO", "Type of grain" and "VALUE" were kept while the rest of the attributes were dropped, the 4 attributes were renamed to ensure consistency with other dimensions. After which, the 6 provinces were concatenated to form one large data frame which consists of the "Volume" of each crop at each month from 2001-2023. The "Farm product prices, crops and livestock" dataset has 15 attributes, "REF_DATE", "GEO", "DGUID", "Farm Products", "UOM", 'UOM_ID", "SCALAR_FACTOR", "SCALAR_ID", "VECTOR", "COORDINATE", "VALUE", "STATUS", "SYMBOL", "TERMINATED", "DECIMALS". Only farm products that consists of the crops within our scope were included. The "REF_DATE", "GEO", "Farm Products" and "VALUE" were kept while the rest of the attributes were dropped, the 4 attributes were renamed to ensure consistency with other dimensions. The

naming convention of the crops were also changed for the two data frames to ensure consistency across the data frames. For example, in the "Volume" data frame 'Wheat, excluding durum': 'Wheat', 'Durum wheat': 'Durum', 'Canola (rapeseed)': 'Canola', whereas in the "Price" data frame, 'Wheat (except durum wheat) [1121111]': 'Wheat', 'Rye [1151152]': 'Rye', 'Oats [115113111]': 'Oats', 'Barley [1151141]': 'Barley'. After ensuring consistency, the two data frames are left joined on the attributes "Date", "Product" and "Province" with "Price" data frame as the left table. Crop_ID, Region_ID, Date_ID and Climate_ID were added to the data frame to form the finalized fact table.

The four dimensions and the fact table were added to a Postgres database hosted on Supabase.

C. Data Quality Issues and Data Integration

Climate Dimension:

This data was obtained from Monthly Climate Summaries provided by Environment Canada containing meteorological data collected from weather stations across Canada. The data is only available on a monthly basis, so to have a complete dataset from January 2001 to December 2023, the individual datasets for each month during that period had to be integrated to produce one complete dataset.

Data Quality Issues:

- Some records had the string '######' as the value for the attributes 'Total_Precip', ''Percent_of_Normal_Precip', or 'Percent_of_Normal_Snowfall'. These were first converted to NaN to have a consistent representation of missing values
- Weather stations that had missing values for a particular attribute were excluded from the calculation of the provincial mean for that attribute in the given month
- Some records had missing values for 'Percent_of_Normal_Snowfall' in the months of July and August. Since 'Snowfall' was 0 for those months, the 'Percent_of_Normal_Snowfall' value was also set to 0 for those records (to be consistent with other complete records where no snowfall fell in months where there is no expected snowfall)
- For some provinces (especially Manitoba, Alberta and Saskatchewan), there were no weather stations in the entire province reporting values for 'Bright_Sunshine_Days' or 'Percent_of_Normal_Bright_Sunshine' for certain months. Since in total this amounted to 58% of

datapoints missing data for at least one of these attributes, it was determined that there were too many missing values to handle, so these two attributes were removed from the Climate dimension

Date Dimension:

There were no data quality issues for this dimension.

Region Dimension:

The geographical data such as area, land and freshwater coverage were not available for download, so the data needed to be manually copied into CSV format to be useable for analysis.

Data Quality Issues:

- Values for population per region were strings so the commas in the values needed to be removed and the values reformatted into integers in order to later perform binning on them.
- Freshwater for PEI was set to '.' where in fact it was 0, so the values needed to be set manually.
- Population data was reported using quarters and years in the columns, so the data needed to be transposed (to be used for duplicate removal) and the date dimension needed to be referenced to get the appropriate effective date given the quarter and year.
- Certain attributes used for classification of regions did not contain a name in French or abbreviations of their name, so the data was kept as NaN for those values in the columns.
- Aggregations for Atlantic, Prairies, and Territories regions needed to be completed manually for area and population values, because of missing raw data.
- Canada has a missing 'NumericCode', so it was manually assigned 0 – a number not used for any other province/territory/region.

Crop Dimension:

There were no data quality issues for this dimension.

Fact Table:

Data Quality Issues:

- Prior to merging, there are NaN values in the "Volume" and "Price" columns in the volume and price data frame respectively. Those rows with both "Volume" and "Price" value as NaN were discarded as we deemed them unable to value add to our project and we will still have more than sufficient data for the later part of the project.
- We used forward fill and backward fill to address the remaining NaN values as we believe there would be some correlation between the current month with the future/prior months.
- Due to the limitation of forward fill and backward fill function, the NaN values at the front and end of the data frame could not be handled appropriately. Therefore, we utilized the global mean values to fill up the 0 values and remaining NaN values.

D. Team Planning

Deliverable checklist	Responsible team member(s)	Expected completion date	Actual completion date	Estimated time (hours) to complete	Actual time (hours) to complete	Notes (if any)
Create database instance	Daniel	March 15th	March 15th	1	1.5	
Create Cimate dimension	Daniel	March 15th	March 15th	1	0.5	

Create Date dimension	Svetlana	March 15th	March 16th	1	1	
Create Region dimension	Svetlana	March 17th	March 17th	4	6	
Create Crop dimension	Yongbin	March 18th	March 18th	3	3	
Staging of dimension Climate	Daniel	March 22nd	March 23rd	3	2	
Staging of dimension Region	Svetlana	March 20th	March 20th	1	1	
Staging of dimension Date	Svetlana	March 20th	March 20th	1	1	
Staging of dimension Crop	Yongbin	March 19th	March 19th	1	1	
Surrogate key pipeline	All	March 19th	March 19th	0.5	0.5	
Staging of fact table – including FKs and measures	Yongbin	March 27th	March 27th	8	8	
Data quality handling and reporting	All	March 22th	March 27th	4	5	

Dimensions

Climate Dimension

```
1 SELECT * FROM public."Climate"
2 ORDER BY "Climate_ID" ASC
```

Data Output | Messages | Notifications

	Climate_ID [PK] bigint	Mean_Temp double precision	Min_Temp double precision	Max_Temp double precision	Mean_Diff_from_Normal double precision	Snowfall double precision	Percent_of_Normal_Snowfall double precision	Total_Precip double precision	Percent_of_Normal_Precip double precision	Snow_on_Ground double precision
1	0	0.351884058	-8.66849711	8.602305476	1.8915	18.86153846	36.08810573	107.7979112	63.11607143	19.377926
2	1	-4.457322176	-18.16610879	8.743096234	6.969306931	4.954146341	16.72413793	5.833195021	22.02521008	6.7222222
3	2	-8.665789474	-24.72903226	4.388311688	6.291262136	6.587407407	34.56363636	7.125	35.32758621	28.449541
4	3	-12.04957265	-29.18632479	2.558974359	5.515873016	10.71282051	43.71604938	10.38529412	45.91860465	41.987179
5	4	-7.284188034	-24.33162393	2.815384615	1.8008	42.97641026	87.45454545	46.65330396	72.24590164	35.673469
6	5	-11.84816514	-27.70412844	-0.005504587	0.778947368	46.93851351	86.67226891	43.58291457	62.79166667	44.892617
7	6	-1.30778098	-13.05444126	9	-1.39798995	15.91420118	95.54424779	47.26276042	52.7699115	18.104234
8	7	-12.2350211	-30.20843882	8.502531646	-3.784693878	16.21746032	92.04504505	15.73033175	91.12389381	12.285714
9	8	-17.14802632	-34.49675325	3.194155844	-5.204901961	11.06015038	91.37614679	10.78141026	87.42608696	32.785046
10	9	-18.76752137	-35.64957265	-1.538461538	-4.688888889	13.37043478	88.35365854	12.84166667	81.62068966	47.301369
11	10	-7.48559322	-22.36737288	7.362711864	0.0168	46.3530303	123.8211382	77.90948276	147.8790323	33.246575
12	11	-11.624	-29.29891304	4.737818182	-0.624375	59.26318408	124.3305785	70.78174603	122.1810345	58
13	12	3.257142857	-7.770845481	13.86938776	0.042929293	15.53802395	56.65625	103.8382199	102.6636771	12.399361
14	13	-1.976470588	-20.55798319	12.65462185	1.792079208	14.15323383	59.68695652	13.95965665	57.22881356	1.8690476
15	14	-4.175974026	-21.70649351	10.59675325	1.281553398	9.274626866	55.12844037	10.11130952	50.45217391	13.379310
16	15	-6.75826087	-23.64869565	5.97826087	0.433870968	19.71932773	113.6790123	21.25211268	93.02325581	29.957746
17	16	-2.883333333	-18.51538462	9.482478632	-0.541269841	37.88542714	140.6048387	42.96137339	76.36	20.985714
18	17	-5.451272727	-27.95309091	8.161454545	-0.596875	59.89004975	164.954955	60.13187251	98.02702703	64.041025

Date Dimension

```
1 SELECT * FROM public."Date"
2 ORDER BY "Date_ID" ASC
```

Data Output | Messages | Notifications

	Date_ID [PK] bigint	Date date	Month_LongName text	Month_ShortName text	Month bigint	NumberOfDaysInMonth bigint	Quarter bigint	NumberOfDaysInQuarter bigint	Year bigint	NumberOfDaysInYear bigint	LeapYear boolean
1	0	2001-01-01	January	Jan	1	31	1	90	2001	365	false
2	1	2001-02-01	February	Feb	2	28	1	90	2001	365	false
3	2	2001-03-01	March	Mar	3	31	1	90	2001	365	false
4	3	2001-04-01	April	Apr	4	30	2	91	2001	365	false
5	4	2001-05-01	May	May	5	31	2	91	2001	365	false
6	5	2001-06-01	June	Jun	6	30	2	91	2001	365	false
7	6	2001-07-01	July	Jul	7	31	3	92	2001	365	false
8	7	2001-08-01	August	Aug	8	31	3	92	2001	365	false
9	8	2001-09-01	September	Sep	9	30	3	92	2001	365	false
10	9	2001-10-01	October	Oct	10	31	4	92	2001	365	false
11	10	2001-11-01	November	Nov	11	30	4	92	2001	365	false
12	11	2001-12-01	December	Dec	12	31	4	92	2001	365	false
13	12	2002-01-01	January	Jan	1	31	1	90	2002	365	false
14	13	2002-02-01	February	Feb	2	28	1	90	2002	365	false
15	14	2002-03-01	March	Mar	3	31	1	90	2002	365	false
16	15	2002-04-01	April	Apr	4	30	2	91	2002	365	false
17	16	2002-05-01	May	May	5	31	2	91	2002	365	false

Region Dimension

```
1 SELECT * FROM public."Region"
2 ORDER BY "Region_ID" ASC
```

Data Output | Messages | Notifications

	Region_ID [PK] bigint	Name_EN text	Name_FR text	Abbrev_EN text	Abbrev_FR text	AlphaCode text	NumericCode bigint	GeographicRegion text	Area_Percent double precision	Area double precision
1	0	Newfoundland and Labrador	Terre-Neuve-et-Labrador	N.L.	T.-N.-L.	NL	10	Atlantic	4.1	4052
2	1	Prince Edward Island	Île-du-Prince-Édouard	P.E.I.	Î.-P.-É.	PE	11	Atlantic	0.1	56
3	2	Nova Scotia	Nouvelle-Écosse	N.S.	N.-É.	NS	12	Atlantic	0.6	552
4	3	Nova Scotia	Nouvelle-Écosse	N.S.	N.-É.	NS	12	Atlantic	0.6	552
5	4	New Brunswick	Nouveau-Brunswick	N.B.	N.-B.	NB	13	Atlantic	0.7	729
6	5	Quebec	Québec	Que.	Qc	QC	24	Canada	15.4	15420
7	6	Ontario	Ontario	Ont.	Ont.	ON	35	Canada	10.8	10763
8	7	Ontario	Ontario	Ont.	Ont.	ON	35	Canada	10.8	10763
9	8	Manitoba	Manitoba	Man.	Man.	MB	46	Prairies	6.5	6477
10	9	Saskatchewan	Saskatchewan	Sask.	Sask.	SK	47	Prairies	6.5	6510
11	10	Saskatchewan	Saskatchewan	Sask.	Sask.	SK	47	Prairies	6.5	6510
12	11	Saskatchewan	Saskatchewan	Sask.	Sask.	SK	47	Prairies	6.5	6510
13	12	Alberta	Alberta	Alta.	Alb.	AB	48	Prairies	6.6	6618
14	13	British Columbia	Colombie-Britannique	B.C.	C.-B.	BC	59	Canada	9.5	9447
15	14	British Columbia	Colombie-Britannique	B.C.	C.-B.	BC	59	Canada	9.5	9447
16	15	Yukon	Yukon	Y.T.	Yn	YT	60	Territories	4.8	4824
17	16	Northwest Territories	Territoires du Nord-Ouest	N.W.T.	T.N.-O.	NT	61	Territories	13.5	13461

Crop Dimension

```
1 SELECT * FROM public."Crop"
2 ORDER BY "Crop_ID" ASC
```

Data Output | Messages | Notifications

	Crop_ID [PK] text	Common Name text	Species text	Family text	Minimum Temperature bigint	Maximum Temperature bigint
1	0	Wheat	Triticum	Grass	5	29
2	1	Durum	Triticum	Grass	5	29
3	2	Oats	Avena sativa	Grass	5	28
4	3	Barley	Hordeum	Grass	5	28
5	4	Rye	Secale	Grass	5	34
6	5	Flaxseed	Linum usitatissimum	Linaceae	5	30
7	6	Canola	Brassica napus	Brassicaceae	3	29

Fact Table

```
1 SELECT * FROM public."FactTable"
2 ORDER BY "Date_ID" ASC, "Crop_ID" ASC, "Climate_ID" ASC, "Region_ID" ASC
```

Data Output | Messages | Notifications

	Date_ID [PK] bigint	Crop_ID [PK] text	Climate_ID [PK] bigint	Region_ID [PK] bigint	Price double precision	Volume double precision
1	0	0	5	5	144.96	97889.0859868666
2	0	2	0	13	98.66	874
3	0	2	1	12	98.66	27873
4	0	2	2	9	90.43	101063
5	0	2	3	8	94.51	81042
6	0	2	4	7	105	1761
7	0	2	5	5	99.19	7229.5
8	0	3	0	13	112.06	4319
9	0	3	1	12	112.06	192621
10	0	3	2	9	88.43	319288
11	0	3	3	8	83.18	59651
12	0	3	4	7	129.03	6585
13	0	3	5	5	131.58	13021
14	0	4	0	13	89.08	3921.5
15	0	4	1	12	89.08	3481