Snowstorm

# Database Design Document

1. Version 0.2
2. 02/23/2019
3. Team: **Ilia Kassianenko, Pawel Kaluski, Fritz Gyger**

Table of Contents

[1. Overview 1](#__RefHeading___Toc9045_565685251)

[2. Assumptions/Constraints/Risks 2](#__RefHeading___Toc9047_565685251)

[2.1 Assumptions 2](#__RefHeading___Toc9049_565685251)

[2.2 Constraints 2](#__RefHeading___Toc9051_565685251)

[2.3 Risks 2](#__RefHeading___Toc9053_565685251)

[3. Design Decisions 3](#__RefHeading___Toc9055_565685251)

[3.1 Key Factors Influencing Design 3](#__RefHeading___Toc9057_565685251)

[3.2 Functional Design Decisions 3](#__RefHeading___Toc9059_565685251)

[3.3 Database Management System Decisions 3](#__RefHeading___Toc9061_565685251)

[3.4 Security and Privacy Design Decisions 3](#__RefHeading___Toc9063_565685251)

[3.5 Performance and Maintenance Design Decisions 3](#__RefHeading___Toc9065_565685251)

[4. Detailed Database Design 5](#__RefHeading___Toc9067_565685251)

[4.1 Roles and Responsibilities 6](#__RefHeading___Toc9069_565685251)

[4.2 Performance Monitoring and Database Efficiency 6](#__RefHeading___Toc9075_565685251)

[4.2.1 Operational Implications 6](#__RefHeading___Toc9077_565685251)

[4.2.2 Data Transfer Requirements 6](#__RefHeading___Toc9079_565685251)

[4.2.3 Data Formats 6](#__RefHeading___Toc9081_565685251)

[Appendix A: Acronyms 7](#__RefHeading___Toc9083_565685251)

List of Figures

No table of figures entries found.

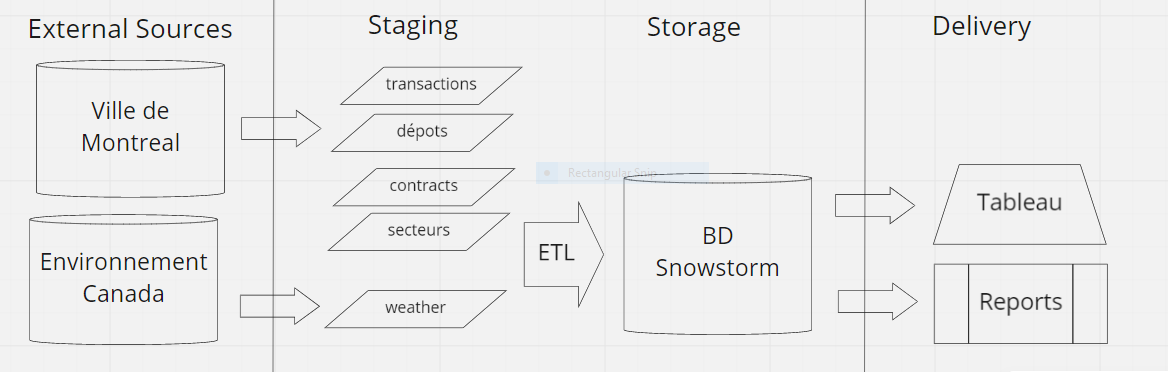
List of Tables

[Table 1 - Acronyms 9](#Table!1|sequence)

## Overview

Instructions: Briefly introduce the system context and the basic design approach or organization, including dependencies on other systems. Identify if the database will supersede or interface with other databases, and specifically identify them if applicable. Also identify interfaces with other systems to the extent that they significantly impact the database design. Discuss the background to the project, if this will help understand the functionality supported by the database design contained in this document.

The purpose of this project is to identify the quantity of snow moved for the City of Montreal to a designated disposal area depot in order to help the city of Montreal as well as the independent snow removers to optimize the snow removal.



We plan to use Alteryx as an ETL tool.

The data is acquired from two main **Data sources:**

**Ville de Montréal** forthe snow removal information **:**

* Transactions déneigement (fichier .csv)

<http://donnees.ville.montreal.qc.ca/dataset/5bfbd75f-7531-48c2-b6b6-072284f7b9e7/resource/dad68871-51b9-4a82-93b0-31cf20b5aa03/download/transactions_deneigement_saison_2018-2019.csv>

* Contrats déneigement (fichier .csv)

<http://donnees.ville.montreal.qc.ca/dataset/5bfbd75f-7531-48c2-b6b6-072284f7b9e7/resource/5dd82872-89f8-439e-9a8a-fff7fea1a28d/download/contrats_deneigement_saison_2018-2019.csv>

* Dépôt de neige (fichier .csv)

<http://donnees.ville.montreal.qc.ca/dataset/8a1d7d54-c297-46fe-b670-bb205641b13e/resource/9ea7b63a-18e1-4e9a-834e-77fd28e55bf8/download/depots_deneigement_saison_2018-2019.csv>

* Secteur de déneigement

<http://donnees.ville.montreal.qc.ca/dataset/9f3911af-3a5f-4c4b-89c7-239ba487b1f1/resource/aa6f2231-9a67-418f-8234-d49462dd6344/download/secteurs_deneigement_saison_2018-2019.csv>

**Environnement Canada** for the weather data form the station at Dorval airport**:**

* Weather data YUL (fichier .csv)

domain: <http://climate.weather.gc.ca/historical_data/search_historic_data_e.html>

<http://climate.weather.gc.ca/climate_data/daily_data_e.html?hlyRange=2008-01-08%7C2019-02-10&dlyRange=2002-12-23%7C2019-02-10&mlyRange=%7C&StationID=30165&Prov=QC&urlExtension=_e.html&searchType=stnName&optLimit=yearRange&StartYear=2018&EndYear=2019&selRowPerPage=25&Line=2&searchMethod=contains&Month=1&Day=1&txtStationName=MONTREAL&timeframe=2&Year=2019>

Each **depot** has a specified capacity, except for a sewage depot, where the runoff sewage systems capacity is not specified. Also, the quantity of daily transactions per depot is considered to be the quantity of visits to unload snow at a specific depot. The daily frequency for a depot is obtained by dividing the quantity of visits for a depot by the total quantity of visits to all depots on a specific day.

Additionally, the amount of precipitation is recorded at the Montreal YUL international airport as an average depth in centimeters. Also, it is assumed for this project that the quantity of precipitation recorded at the Montreal YUL international airport is treated as the uniform distribution of precipitation across the area, which is being consider for this analysis. Therefore, all locations serviced by vehicles with a **contract** ID are assumed to have received the same amount of precipitation as the aforementioned airport.

Each daily snow volume unloaded at each depot is divided by the recorded daily average depth of the precipitation. The result from the division is an order of magnitude estimate for the daily surface area that has received precipitation and that has been served by the fleet of vehicles.

Also, by dividing the quantity of visits by the surface area where snow was removed, a service ratio is obtained for a specific depot and day. The difference in time for a transaction between the snow loading and unloading time is considered to be the cycle time. The project herein does not consider the effect on the cycle time following re-assignment of vehicles to depots.

Finally, this analysis will produce for each depot the daily frequency of use and the daily service ratio. Consequently, the analysis will show if constrained capacity depots are used more frequently daily than a sewer depot, and the analysis will show at which depot there is an high quantity of vehicles unloading snow.

The data can also be used to compare the different boroughs and their sectors or the different contracts. For example which contract delivers the best value?

For the **reporting** we plan to use Tableau. We may give external customer access for a small fee.

Since our system is informational (vs operational) we use the **BASE property** for the project:

Basic availability - Allows systems to be temporarily inconsistent so that transactions are manageable.

Soft-state - Recognizes that some inaccuracy is temporarily allowed and data may change while being used to reduce the amount of consumed resources.

Eventual consistency – As the name suggests means eventually, when all service logic is executed, the system is left in a consistent state.

+ Ilia (via Slack)

Licence: Attribution 4.0 international CC BY 4.0

*Project requirements concerning the ETL assumptions: how data is extracted, transformed, and loaded into the db ? Or should we make assumptions and focus primarily on the software and hardware resources of the db ?*

Daniel [12:57 PM]

*I want to have a high level of both. You don't need to go in details, but simply mention why the option was chosen. Ex. "I chose Spark due to be able to process real-time and batch streamings with the same code." But you don't need to give the specifications of how the clusters will be configured.*

## Assumptions/Constraints/Risks

### Assumptions

Instructions: Describe any assumptions or dependencies regarding the database design for the system. These may concern such issues as: related software or hardware, operating systems, or end-user characteristics.

### Constraints

Instructions: Describe any limitations or constraints that have a significant impact on the database design for the system.

* Weekly extractions as the files are updated weekly by the city, every Monday at 1am EST between November 1st – April 1st. (weather Canada tbc - Ilia?)

### Risks

Instructions: Describe any risks associated with the database design and proposed mitigation strategies.

Changes of source URL, source format

## Design Decisions

Instructions: Utilizing the following subsections, describe decisions made that impact the proposed database design. This should include the platform and database management system (DBMS) chosen for the project. Include any other information relevant to the database design decisions (e.g., Data Conversion Plan, Service Level Agreements (SLAs)). The Design Decisions section is written at a higher level than the subsequent Detailed Database Design section, and provides an understanding and rationale for the content in the Detailed Database Design section.

### Key Factors Influencing Design

Instructions: Describe key functional or non-functional requirements that influenced the design. If all such decisions are explicit in the requirements, this section shall so state. Design decisions that respond to requirements designated as critical (e.g., those for performance, availability, security, or privacy) shall be placed in separate subparagraphs.

### Functional Design Decisions

Instructions: Describe decisions about how the database will behave in meeting its requirements from a user's point of view (i.e., functionality of the database from an application perspective), ignoring internal implementation, and any other decisions affecting further design of the database. Include decisions regarding inputs the database will accept and outputs (displays, reports, messages, responses, etc.) it will need to support, including interfaces with other systems. Describe the general types of processing (sequential versus random for inserts, updates, deletes and queries) required both for data entering the database, and data most frequently accessed. Also include decisions on how databases/data files will appear to the user.

### Database Management System Decisions

Instructions: Describe design decisions regarding the DBMS intended for the initial implementation. Provide the name of the DBMS, the reason for selection, and the type of flexibility built into the database for adapting to changing requirements.

### Security and Privacy Design Decisions

Instructions: Describe design decisions on the levels and types of security and privacy to be offered by the database. General descriptions of classifications of users and their general access rights should be included.

No confidential or private data, all data retrieved is publicly available on the web. \*

No IP from outside Canada is allowed to use the system.

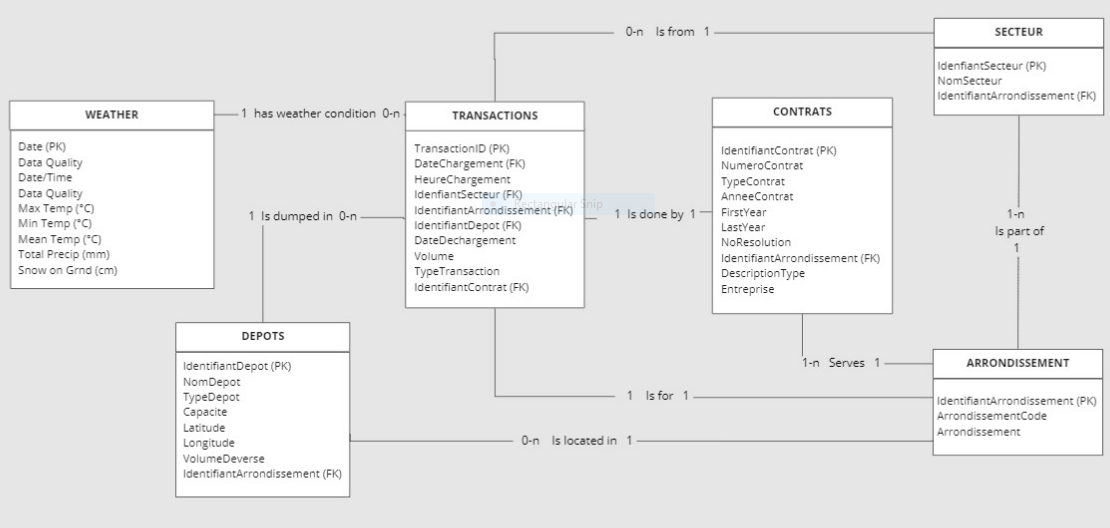
### Performance and Maintenance Design Decisions

Instructions: Describe how performance and availability requirements will be met. Examples include:

* Describe design decisions on database distribution (such as client/server), master database file updates and maintenance, including maintaining consistency, establishing/ reestablishing and maintaining synchronization, enforcing integrity and business rules.
* Describe design decisions to address concurrence issues (e.g., how the data are partitioned or distributed to support multiple applications or competing update functions, if applicable).
* Describe design decisions to support Service Level Agreements (SLAs) for key functions supported by the database.
* Describe design decisions on backup and restoration including data and process distribution strategies, permissible actions during backup and restoration, and special considerations for new or non-standard technologies such as video and sound. Describe the impact this maintenance will have on availability.
* Describe design decisions on data reorganization (i.e., repacking, sorting, table and index maintenance), synchronization, and consistency, including automated disk management and space reclamation considerations, optimizing strategies and considerations, storage and size considerations (e.g., future expansion), and population of the database and capture of legacy data. Describe the impact this maintenance will have on availability.
* Describe design decisions to support purging and/or archiving of data to ensure performance and storage objectives are met. Describe the impact this maintenance will have on availability. Describe any needs to recall archived data back into the database.

## Detailed Database Design

Instructions: Describe the design of all DBMS structure associated with the system. The headings and sub-headings in this section should be structured according to the information to be presented, and may include discussions about or references to the following:

* Conceptual Data Model (CDM)
* Logical Data Model (LDM) and LDM Entity Relationship Diagram (ERD).
* 

<https://realtimeboard.com/app/board/o9J_kyQw7mY=/?userEmail=fgyger@gmail.com&invite=2aaaaaaadcc970d363df49da3d893532-f09c55633fb9bfe1-e627fdad19a479bc-f83312ccebd6dc98&event=mailInvite&mailUserEmail=fgyger@gmail.com&track=true%22>

* Physical Data Model (PDM) with a description of the DBMS schemas, sub-schemas, records, sets, tables.
* A comprehensive Data Dictionary showing data stores, data element name, type, length, source, constraints, validation rules, maintenance (create, read, update, delete (CRUD) capability), audit and data masking requirements, expected data volumes, life expectancy of the data, information life-cycle management strategy or at least an archiving strategy, outputs, aliases, and description.
* Planned implementation factors (e.g., distribution and synchronization) that impact the design.
* Estimate of the DBMS file size or volume of data per entity.
* Definition of the update frequency of the database tables, views, files, areas, records, sets, and data pages. Also provide an estimate of the number of transactions, if the database is an online transaction-based system.

The detailed database design information can be included as an appendix, such as DDLs, which would be referenced here.

### Roles and Responsibilities

Instructions: Identify the organizations and personnel responsible for the following database administrative functions: database administrator, system administrator, and security administrator. Describe specific administration skill requirements applicable to the database.

DBA : SQL,

Systems admin :

Security admin : n/a

Data Acquisition & cleansing : ETL

### Performance Monitoring and Database Efficiency

#### Operational Implications

Instructions: Describe operational implications of data transfer, refresh and update scenarios and expected windows.

* Transactions & Weather :
  + Weekly Mondays after 1am : the city updates the files every Monday at 1am EST
  + Weather TBC

#### Data Transfer Requirements

Instructions: Describe data transfer requirements to and from the software, including data content, format, sequence, volume/frequency and any conversion issues.

* Extract source files in UTF-8 format in order to have French accents (ex. Saint-Léonard instead of Saint-LÃ©onard)
* Volume :
  + Initial : about 35 MB
  + Weekly growth : about 10MB for transaction & weather data.
* Data content and format (See 4.2.3 Data Formats)
* Data Transfer?

#### Data Formats / Data Disctionnary

Instructions: Describe formats of data for both the sending and receiving systems, including the data item names, codes, or abbreviations that are to be interchanged, as well as any units of measure/conversion issues.

**Source files**

**depots**\_deneigement\_saison\_2018-2019, .csv files

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Type** | **Length** | **Nullable** | **Comments** |
| IdentifiantDepot | INT |  | N |  |
| NomDepot | VARCHAR | 30 | N |  |
| TypeDepot | VARCHAR | 15 | N | Values : Carierre, Chute\_Egout, Entassement, |
| Capacite | INT |  | Y | 0 = Sewer well |
| Latitude | DECIMAL | 8,6 | N |  |
| Longitude | DECIMAL | 8,6 | N |  |
| VolumeDeverse | DECIMAL | 9,2 | N |  |
| IdentifiantArrondissement | INT |  | N |  |
| ArrondissementCode | CHAR | 3 | N |  |
| Arrondissement | VARCHAR | 30 | N |  |
| MTM8\_X | DECIMAL | 7,1 | N | ?? |
| MTM8\_Y | DECIMAL | 8,1 | N | ?? |

**secteurs**\_deneigement\_saison\_2018-2019, csv file

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Type** | **Length** | **Nullable** | **Comments** |
| IdenfiantSecteur | INT |  | N |  |
| NomSecteur | VARCHAR | 7 | N |  |
| IdentifiantArrondissement | INT |  | N |  |
| ArrondissementCode | CHAR | 3 | N |  |
| Arrondissement | VARCHAR | 30 | N |  |

**Data dictionary:** <http://donnees.ville.montreal.qc.ca/dataset/contrats-transaction-deneigement>

**contrats**\_deneigement\_saison\_2018-2019, csv file

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Type** | **Length** | **Nullable** | **Comments** |
| IdentifiantContrat | INT |  | N |  |
| NumeroContrat | VARCHAR | 7 | N |  |
| TypeContrat | VARCHAR | 12 | Y | Values : Déneigement, Régie, Transport, blank |
| AnneeContrat | CHAR | 9 | N |  |
| NoResolution | VARCHAR | 11 | Y |  |
| IdentifiantArrondissement | INT |  | N |  |
| ArrondissementCode | CHAR | 3 | N |  |
| Arrondissement | VARCHAR | 30 | N |  |
| DescriptionType | VARCHAR | 100 | Y |  |
| Entreprise | VARCHAR | 50 | N |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

**transactions**\_deneigement\_saison\_2018-2019, csv file

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Type** | **Length** | **Nullable** | **Comments** |
| TransactionID | INT |  | N |  |
| DateChargement | TIMESTAMP | 19 | N | YYYY-MM-DD hh:mm:ss |
| IdenfiantSecteur | INT |  | Y |  |
| NomSecteur | VARCHAR | 7 | Y |  |
| IdentifiantArrondissement | INT |  | N |  |
| ArrondissementCode | CHAR | 3 | N |  |
| Arrondissement | VARCHAR | 30 | N |  |
| IdentifiantDepot | INT |  | N |  |
| NomDepot | VARCHAR | 30 | N |  |
| DateDechargement | TIMESTAMP | 19 | N |  |
| Volume | DECIMAL | 4,2 | N |  |
| TypeTransaction | VARCHAR | 7 | N | Values :  AUT, N-AUT, BARCODE |
| IdentifiantContrat | INT |  | Y |  |
| NumeroContrat | VARCHAR | 7 | Y |  |

**weather\_data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Type** | **Length** | **Nullable** | **Comments** |
| Date/Time | DATE | 10 | N | YYYY-MM-DD |
| Year | INT | 4 | N |  |
| Month | INT | 2 | N |  |
| Day | INT | 2 | N |  |
| Data Quality | ?? | ?? | Y |  |
| Max Temp | DECIMAL | 3,1 | Y |  |
| Max Temp Flag | ?? | ?? | Y |  |
| Min Temp (°C) | DECIMAL | 3,1 | Y |  |
| Min Temp Flag | ?? | ?? | Y |  |
| Mean Temp (°C) | DECIMAL | 3,1 | Y |  |
| Mean Temp Flag | ?? | ?? | Y |  |
| Heat Deg Days (°C) | DECIMAL | 3,1 | Y |  |
| Heat Deg Days Flag | ?? | ?? | Y |  |
| Cool Deg Days (°C) | DECIMAL | 3,1 | Y | Always 0 |
| Cool Deg Days Flag | ?? | ?? | Y | Always blank |
| Total Rain (mm) | ?? | ?? | Y |  |
| Total Rain Flag | ?? | ?? | Y |  |
| Total Snow (cm) | ?? | ?? | Y |  |
| Total Snow Flag | CHAR | 1 | Y | Value : M |
| Total Precip (mm) | DECIMAL | 3,1 | Y |  |
| Total Precip Flag | CHAR | 1 | Y | Value : M |
| Snow on Grnd (cm) | INT | 2 | Y |  |
| Snow on Grnd Flag | ?? | ?? | Y | Always blank |
| Dir of Max Gust (10s deg) | INT | 2 | Y |  |
| Dir of Max Gust Flag | CHAR | 1 | Y | Value : M |
| Spd of Max Gust (km/h) | INT | 2 | Y |  |
| Spd of Max Gust Flag | CHAR | 1 | Y | Value : M |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |

**Validation rules:**

* DATES : Need to be valid dates in the right sequence

(ex. 2106-02-07 found in source data)

**Transformation rules:**

* TRANSACTIONS
  + Split DateChargement (format YYYY-MM-DD hh:mm:ss) into   
    DateChargement (YYYY-MM-DD) and HeureChargement (hh:mm:ss) to allow easy join with DATE (YYYY-MM-DD) in WEATHER table.
* CONTRATS
  + Split AnneeContrat (format YYYY-YYYY) in FirstYear & LastYear in order to allow easier filtering.

**Target tables**

**DEPOTS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Type** | **Length** | **Nullable** | **Comments** |
| **IdentifiantDepot** | INT |  | N | Primary Key |
| NomDepot | VARCHAR | 30 | N |  |
| TypeDepot | VARCHAR | 15 | N | Values : Carierre, Chute\_Egout, Entassement, |
| Capacite | INT |  | Y | 0 = Sewer well |
| Latitude | DECIMAL | 8,6 | N |  |
| Longitude | DECIMAL | 8,6 | N |  |
| VolumeDeverse | DECIMAL | 9,2 | N |  |
| IdentifiantArrondissement | INT |  | N | Foreign Key |

**SECTEUR**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Type** | **Length** | **Nullable** | **Comments** |
| **IdenfiantSecteur** | INT |  | N | Primary Key |
| NomSecteur | VARCHAR | 7 | N |  |
| IdentifiantArrondissement | INT |  | N | Foreign Key |

**ARRONDISSEMENT**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Type** | **Length** | **Nullable** | **Comments** |
| IdentifiantArrondissement | INT |  | N | Primary Key |
| ArrondissementCode | CHAR | 3 | N |  |
| Arrondissement | VARCHAR | 30 | N |  |

**CONTRATS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Type** | **Length** | **Nullable** | **Comments** |
| **IdentifiantContrat** | INT |  | N | Primary Key |
| NumeroContrat | VARCHAR | 7 | N |  |
| TypeContrat | VARCHAR | 12 | Y | Values : Déneigement, Régie, Transport, blank |
| AnneeContrat | CHAR | 9 | N |  |
| FirstYear | CHAR | 4 | N |  |
| LastYear | CHAR | 4 | N |  |
| NoResolution | VARCHAR | 11 | Y |  |
| IdentifiantArrondissement | INT |  | N | Foreign Key |
| DescriptionType | VARCHAR | 100 | Y |  |
| Entreprise | VARCHAR | 50 | N |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |

**TRANSACTIONS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Type** | **Length** | **Nullable** | **Comments** |
| **TransactionID** | INT |  | N | Primary Key |
| DateChargement | DATE | 10 | N | Foreign Key  YYYY-MM-DD |
| TimeChargement | TIME | 8 | N | hh:mm:ss |
| IdenfiantSecteur | INT |  | Y | Foreign Key |
| IdentifiantArrondissement | INT |  | N | Foreign Key |
| IdentifiantDepot | INT |  | N | Foreign Key |
| DateDechargement | TIMESTAMP | 19 | N |  |
| Volume | DECIMAL | 4,2 | N |  |
| TypeTransaction | VARCHAR | 7 | N | Values :  AUT, N-AUT, BARCODE |
| IdentifiantContrat | INT |  | Y | Foreign Key |

**WEATHER**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Type** | **Length** | **Nullable** | **Comments** |
| **Date** | DATE | 10 | N | Primary Key  YYYY-MM-DD |
| Data Quality | ?? | ?? | Y |  |
| Max Temp | DECIMAL | 3,1 | Y | Celsius |
| Min Temp | DECIMAL | 3,1 | Y | Celsius |
| Mean Temp | DECIMAL | 3,1 | Y | Celsius |
| Mean Temp Flag | ?? | ?? | Y |  |
| Total Precip) | DECIMAL | 3,1 | Y | Millimeters |
| Snow on Grnd | INT | 2 | Y | Centimeters |

DDL Ilia & Pavel

Appendix A: Acronyms

Instructions: Provide a list of acronyms and associated literal translations used within the document. List the acronyms in alphabetical order using a tabular format as depicted below.

Table 1 - Acronyms

| Acronym | Literal Translation |
| --- | --- |
| csv | Comma Separated Values |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

DDLs

