Snowstorm

# Database Design Document

1. Version 0.8
2. 03/19/2019
3. Team: **Ilia Kassianenko, Pawel Kaluski, Fritz Gyger**

****Instructor:** Daniel Paes**

**CEBD-1250 – Big Data Storage**

Table of Contents

[1. Overview 4](#_Toc4005036)

[1.1 Scope 4](#_Toc4005037)

[1.2 Existing applications 4](#_Toc4005038)

[1.3 Data acquisition 4](#_Toc4005039)

[1.4 Contexts for Analytics 5](#_Toc4005040)

[1.5 Timeframe 5](#_Toc4005041)

[1.6 Base Property for Data 5](#_Toc4005042)

[2. Assumptions/Constraints/Risks 6](#_Toc4005043)

[2.1 Assumptions 6](#_Toc4005044)

[2.1.1 Ville de Montréal 6](#_Toc4005047)

[2.1.2 Environment Canada 6](#_Toc4005050)

[2.2 Constraints 7](#_Toc4005052)

[2.3 Risks 7](#_Toc4005053)

[3. Design Decisions 8](#_Toc4005054)

[3.1 Key Factors Influencing Design 8](#_Toc4005062)

[3.2 Functional Design Decisions 8](#_Toc4005063)

[3.3 Database Management System Decisions 9](#_Toc4005064)

[3.4 Security and Privacy Design Decisions 9](#_Toc4005066)

[3.5 Performance and Maintenance Design Decisions 9](#_Toc4005067)

[4. Detailed Database Design 11](#_Toc4005068)

[4.1 Roles and Responsibilities 14](#_Toc4005069)

[4.2 Performance Monitoring and Database Efficiency 14](#_Toc4005070)

[4.2.1 Operational Implications 14](#_Toc4005071)

[4.2.2 Data Transfer Requirements 14](#_Toc4005072)

[4.2.3 Data Formats / Data Dictionary 16](#_Toc4005074)

[5. Appendix A: Acronyms 25](#_Toc4005097)

[6. Appendix B: DDL 26](#_Toc4005098)

[7. Appendix C: Decision chart database system 27](#_Toc4005099)

[8. Appendix D: Decision tree 28](#_Toc4005100)

[9. Appendix E: Highlight of milestone meetings 29](#_Toc4005101)

[10. Appendix F: Decision Chart Storage 30](#_Toc4005102)

[11. References 31](#_Toc4005103)

## Overview

Within the region defined by the City of Montreal, “the average volume of snow taken to disposal sites per year is 12 million cubic meters – 300,000 truckloads.” [1\*]

### Scope

This proposal includes the database architecture for application, which enables the identification of the quantity of snow, which is moved to designated disposal area depots [4\*] that are operated by the City of Montreal (COM). Also, the application includes the establishment of performance metrics for the third parties, which were awarded snow removal contracts by the COM.

The primary target users are COM official and residents. Secondary users are third parties, which plan on bidding on snow removal contracts to the COM and which are interested gaging their relative performance.

### Existing applications

This application is complementary to existing COM applications. Specifically, “since the winter of 2014, Montréal has used specialized software that feeds the INFO-Neige application (available on the AppleStore and GooglePlay) using open data.” [1\*] Planif-Neige is used by boroughs for planning the path of snow loading vehicles and it indicate the progress of the operations; Planif-Neige feeds INFO-Neige and the Web snow removal map with open data. [2\*]

Additionally, “since the winter of 2014, Montréal has used specialized management software [SIT-Neige (Système Intelligent de Transport de la Neige)] to optimize snow removal operations (loading, transportation and elimination) and monitor billing.”[1\*] SIT-Neige includes a system and telemetry equipment, which is used to manage the transactions related to snow loading and removal; GPS equipment allows to feed the location of snow blowers to INFO-Neige and snow removal map applications. [2\*]

### Data acquisition

There are two online sources for data acquisitions: Ville de Montreal (COM) and Environment Canada (EC). Moreover, municipal snow removal data is provided by COM, and weather data is provided by Environment Canada.

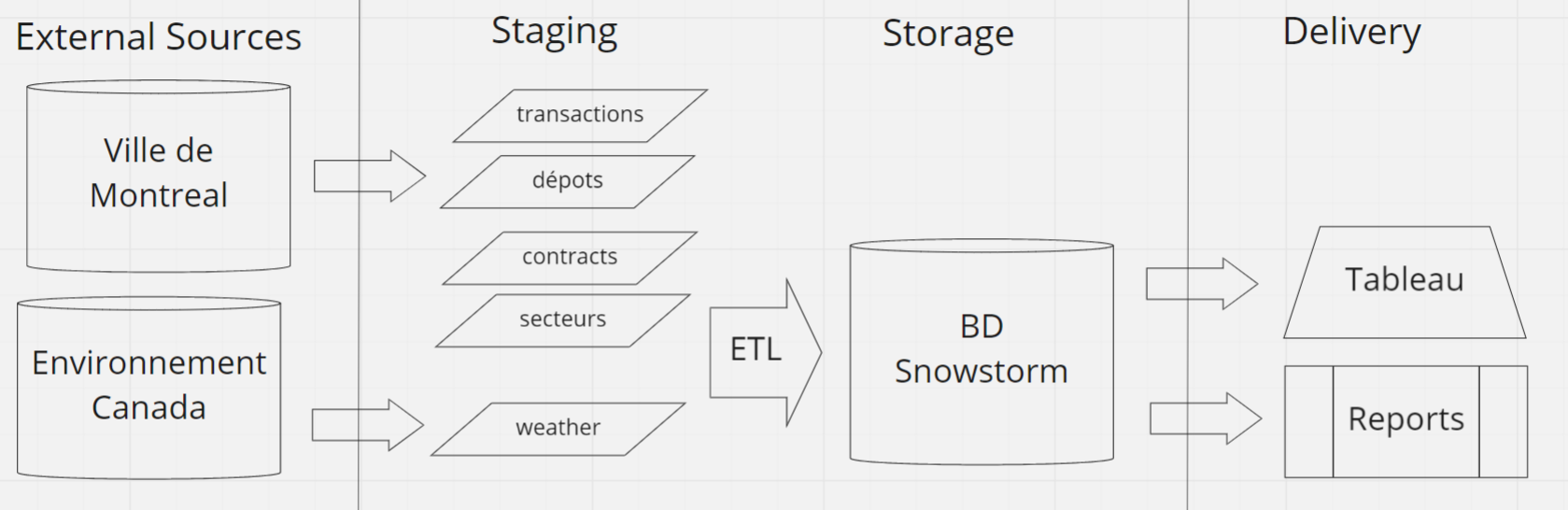


Figure 1 Data Sources

The details concerning these data sources are further discussed in sections [2.1. Assumptions](#_Assumptions) and [4.2.3 Data Formats / Data Dictionary](#__RefHeading___Toc9081_565685251).

### Contexts for Analytics

As defined in the project proposal submitted for CEBD-1160, 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.

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 depots, boroughs and their sectors or the different contracts. Finally, it will be possible to measure relative performance between third parties, which were awarded snow removal contracts.

### Timeframe

We plan to keep 10 years of history in order to compare different year and are able to see trends.

### Base Property for Data

Our system is informational, rather than operational; we propose to use the BASE property of data for the project:

Specifically, there is an expectation that the CSV will be consistently available every Monday at 1 AM EST. Also, the COM provides a disclaimer that previously released records may be corrected retroactively.

Service failure is not critical within our environment, since the COM holds all records. If the portal fails or records become unavailable, the native files are assumed to be available from city of Montreal snow removal management system, in reference to the disclaimer provided by the COM.

One transaction and one contract file are downloaded once a week, and one weather report is obtained once a week. There are no other consistent inputs, and deports quantity and properties are assumed to remain unchanged.

*Data may overwritten when the CSV files are downloaded from the city of Montreal open data portal. ???*

## 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.

There are two online sources for data acquisitions: Ville de Montreal (COM) and Environment Canada.

#### 

A portal by Ville de Montréal (COM) is accessed data, which is available through an license. The data is available through four distinct reports.

* [1]
* [2]
* [3]
* [4]

#### Environment Canada

A portal for is access. Specifically, a filter is applied in order to obtain data from weather station (YUL or “PIERRE ELLIOTT TRUDEAU INTL”).

* [5] [6]

T

Additionally, the amount of precipitation is recorded at the Montreal YUL international airport as an average depth in centimeters

The source URL for each file is dependent on the chosen time period for which records are exported, and it is not unique. The file is exported manually from the EC portal.

### Constraints

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

* For the 2018-2019 season, by using the same source URL, automated weekly extractions are possible, as the files are updated weekly by the COM, every Monday at 1am EST between November 1st – April 1st.
* Extraction of weather data from Environment Canada is performed manually every Monday at 1am EST between November 1st – April 1st. This method is used until an automated method is determined to be possible.

### Risks

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

* The COM can change of the source URL and the file format without providing notice to its users.
* EC can also make changes to the file format without providing notice to its users.
* COM provides a disclaimer that previously released records may be corrected retroactively [3\*]. Previously, reported and analyzed data may overwrite.
* Digital Ocean is a privately funded entity, which may include Venture Capital as a funding source, as shown with prior rounds of funding are documented on Crunchbase [3\*]. A privately funded entity is not mandated to publish publicly its financial statements.
  + The solvency of this entity cannot be assessed and a credit scope is not publicly available. In case of its insolvency, the entity may not provide notice to its users.
  + In case of systemic insolvency, which includes other entities operating within the same sector, a possible alternative may not be available for individual use and within the same magnitude for price.

## 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

Based on pricing and features shown on Capterra [7\*], Digital Ocean [8\*], and Cloudways [9\*], different providers were compared. Pricing was the primary factor and the availability of feature to meet our needs was the secondary factor, for proposing the chosen solution.

### 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.

For the proposed data architecture, structured data with a low volume will be used. Where files size is measured in MB. Also, due to the weekly frequency of updates and due to the aggregation use case, we propose the use of a relational database with a warm storage option.

See also Appendix C & D.

See 3.5

### 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**.

Inputs files in “csv” format, which are obtained from different sources and loaded in weekly, by using batch processes on Monday EDT mornings.

We plan to use Python to develop the ETL process and the data cleansing.

See rules in section 4.3.

For the **reporting** we plan to use Tableau to develop reports and dashboards. We may give external customer access for a small fee (tbd) for self-service.

Currently there are no interfaces with other systems planned, but could be added later.

### 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.

The proposed solution, Digital Ocean, is considered to offer the need functionality at the lowest price. Should there be a change of magnitude to the price, the proposed alternative is Caspio.

information, whichresulted in this assessment, are shown within

Storage options: [7] <https://www.digitalocean.com/products/managed-databases/> Pawel

Digital Ocean offers PostgreSQL, one of the most advanced SQL-compliant, open source object-relational database system. They handle the maintenance operations and updates with free technical support 24/7.

The database can be created, accessed and managed either through their User Interface or through their API.

We can monitor your cluster with minute-by-minute performance graphs and upgrade our nodes with horizontal read scaling as we grow.

### 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.

Publicly accessible data from COM and EC is used, and user accounts are not used. Only the accounts of team members with access to the Digital Ocean account need to be secured. Network security and hardware security are managed by the solution provider, Digital Ocean.

A two-factor authentication process is not in place for Digital Ocean.

Team members will be prompted by our Security Administrator every three months to reset their password.

A method for IP filtering will be researched, in order to enable access to only visitors with a Canadian IP address, which are the target users for this information.

With Digital Ocean, it is possible user an encryption file system on an SSD block storage volume with Linux Unified Key Setup (LUKS). [10\*]

An SSL (Secure Sockets Layer) certificate will be needed to secure the website, where our application results will be shown. An SSL certificate is used on the Digital Ocean webstire.

### 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).

#### Service Level Agreements

* Describe design decisions to support Service Level Agreements (SLAs) for key functions supported by the database.
* As target users reside within Canada’s time zones, the following Service Level Agreement (SLAs) will be used.
* During Eastern Daylight Time (EDT) zone hours and during week days, from 9 AM – 5PM, members of the team can be reached by online visitors by e-mail.
* Application failure is first investigated each week day.
* Normal site operations will occur during week days, between 8 AM – 8 PM EDT, which enables to reach visitors on the eastern and western coasts.
* Outside of normal site operations hours, only downloadable content is accessible, and support tickets can be created with an integration to Spiceworks [5\*].
* Support ticket resolution will occur the following week day.
* This permits downtime for operations and maintenance, for archiving purposes and for automated ETL processes for the source files.

#### Backup and restoration

* 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.

Backups and restorations are handled by Digital Ocean

Daniel: In case of a structured database which will be the backup policy in place. = depends on Provider we’ll chose (Pawel?)

Data is backed up every day by Digital Ocean and can be restored to any point within the previous seven days. Since the data is available going back several years at the source systems, we’ll download and load it again in case we need to restore data further back.

Digital Ocean uses RAM on the virtual machine and SSD for storage. [7] With weekly updates our data is considered warm.

Since we have low volume and plan to keep 10 years history, no data need to be archived. It is also available at the source systems.

We will start with a single node managed database cluster with 1GB of RAM. They feature automatic failback with standby nodes, so if a node fails data handling is automatically switched to a standby node.

Each cluster allows 25 backend connections per 1 GB of RAM minus 3 connections per node that are reserved for maintenance, which means 22 available. These will be sufficient as a start, later we can look into connection pooling.

We can select date and time during which weekly automatic updates for the database engine and operating system will occur to keep the service stable and secure. We’ll schedule these updates during the night.

* 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.

Digital Ocean, uses RAM on the virtual machine and SSD for storage. With weekly updates our data is considered warm.

Daniel : "Describe design decisions on data reorganization" is if the data will be stored as **SSD** or HDD, RAM, data heat (hot **warm** cold), and if you guys have some policies on moving data from "hot" to "warm" type of storage, is it expected to have a stop on the application to do so?

* 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.

Since we have low volume and plan to keep 10 years history, no data need to be archived.

## 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:

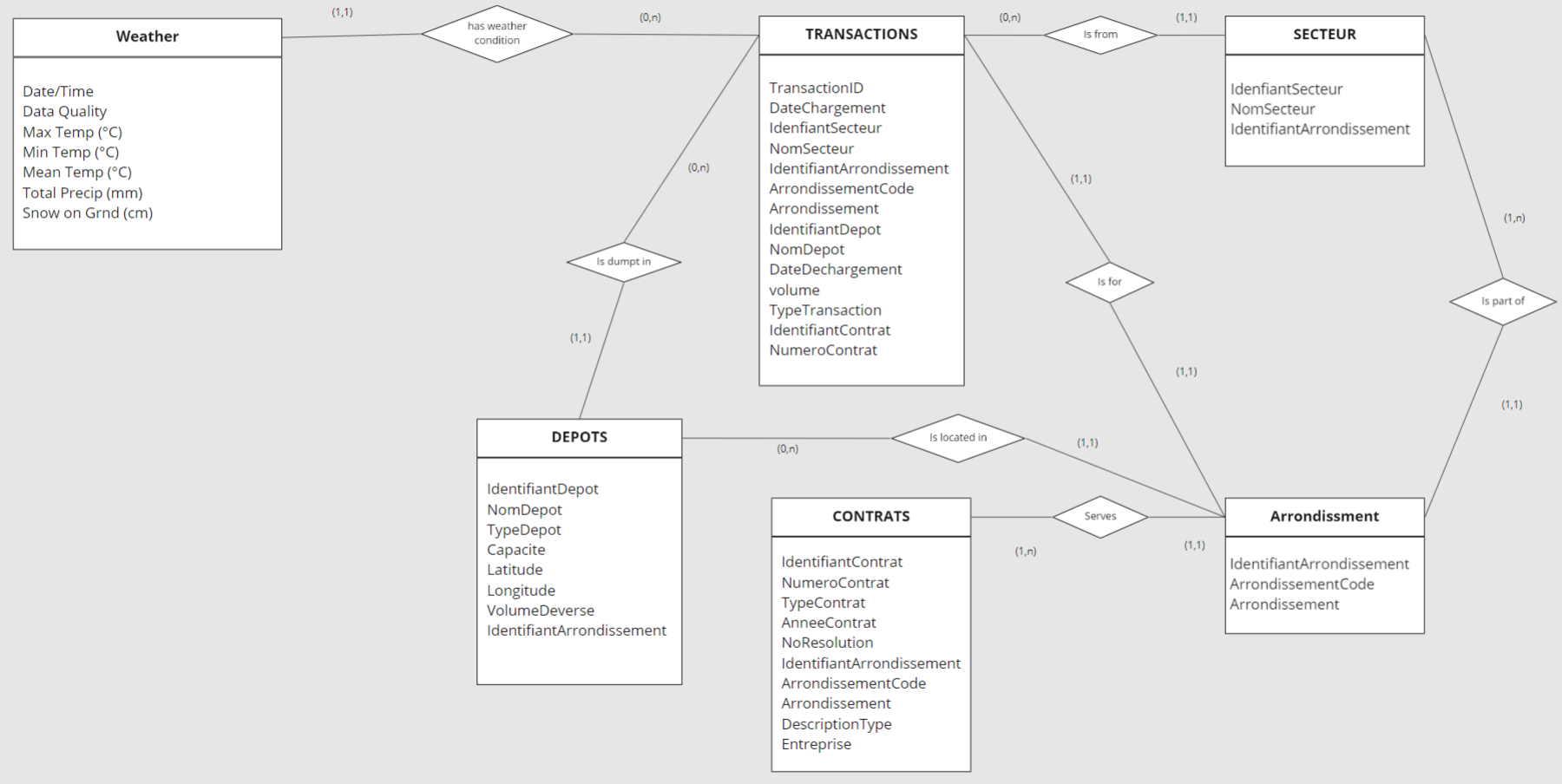


Figure 2 Conceptual Data Model (CDM)

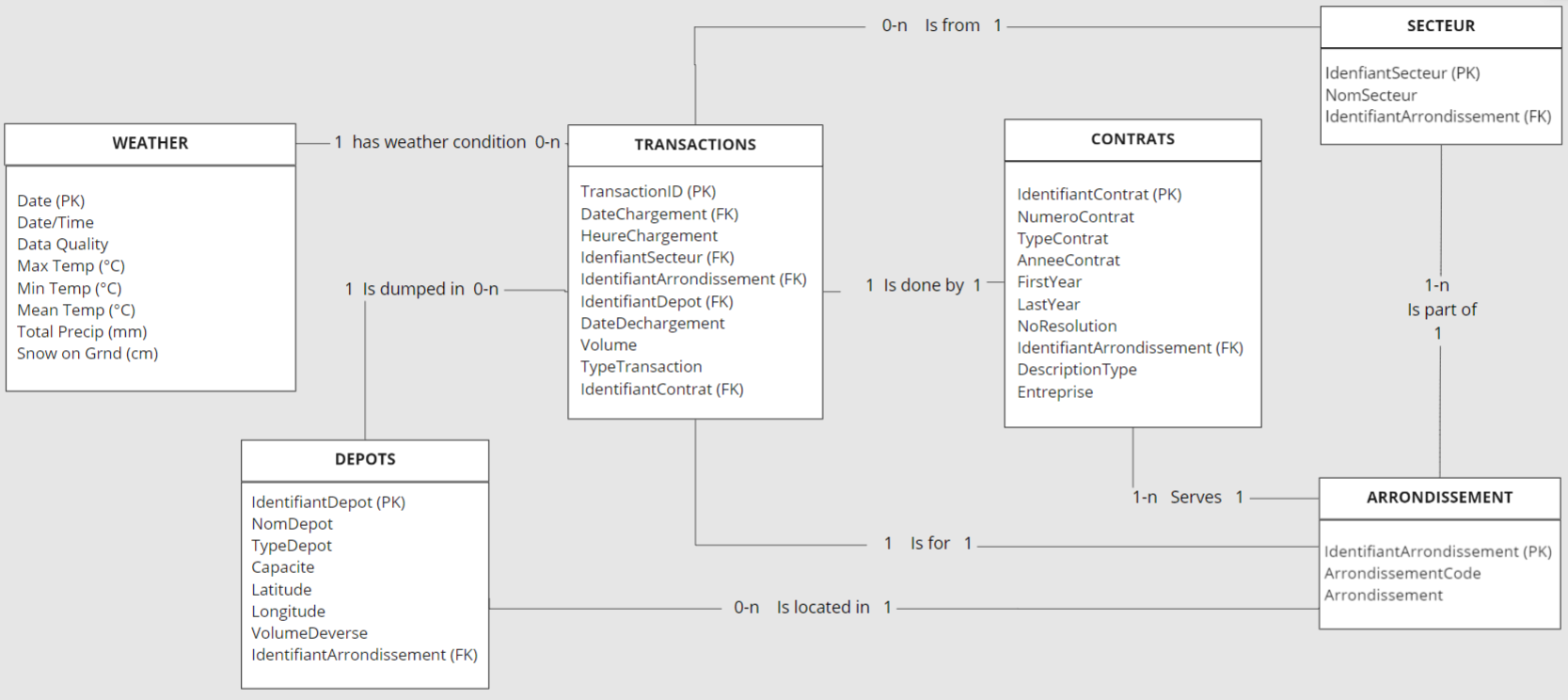


Figure 3 Logical Data Model (LDM) and LDM Entity Relationship Diagram (ERD).

[8] <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.

See Appendix B DDLs

* 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.

CRUD – text to be added

* Planned implementation factors (e.g., distribution and synchronization) that impact the design.
* Estimate of the DBMS file size or volume of data per entity.
* Volume :
  + Initial : about 35 MB
  + Weekly growth: about 10MB for transaction & weather data.
* 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.
* Transactions & Weather :
  + Weekly Mondays after 1am : the city updates the files every Monday at 1am EST
  + Weather Mondays after 1am

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

Daniel : At 4.1 is simply listing the departments which will be responsible to manage the system. for example "The database administration will be handled by the Infrastructure department, where as "Query optimization and error handling" will be handled by production support.

### 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 Mondays after 1am

#### Data Transfer Requirements

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

There are two online sources for data acquisitions: Ville de Montreal (COM) and Environment Canada.

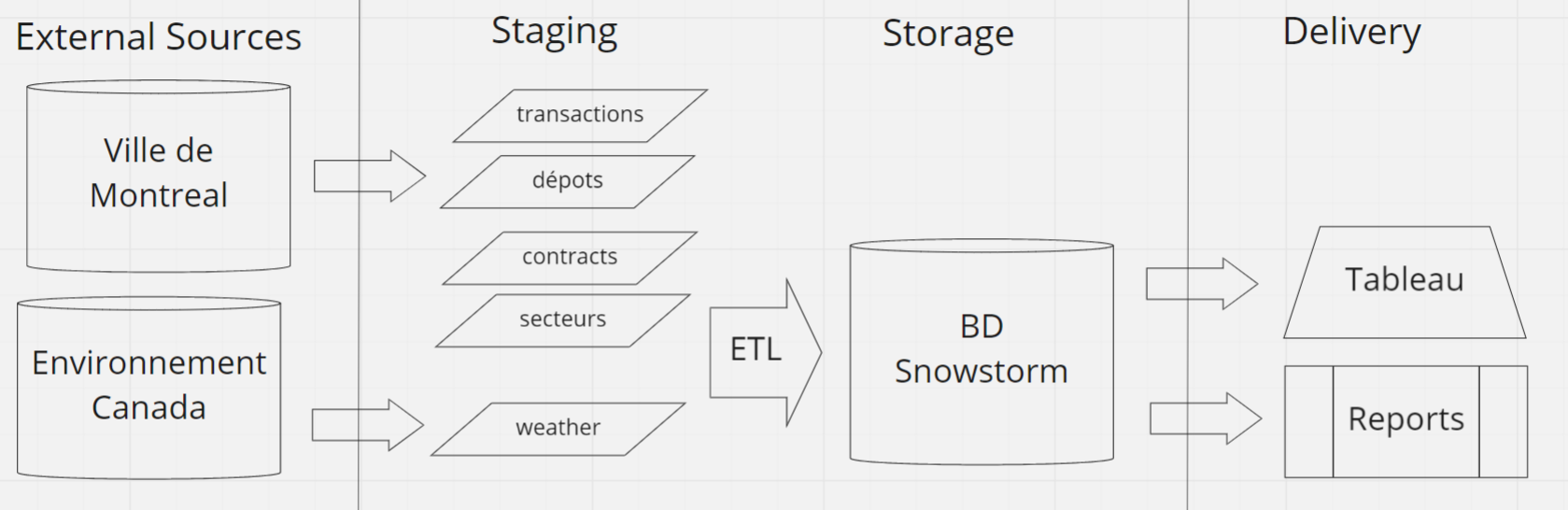


Figure 4 Data Sources

* 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_Formats_/))

#### Data Formats / Data Dictionary

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.

##### Data Formats / Data Dictionary

##### Source files

Table 1 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 | A “0” value corresponds to a 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 | Not defined and no response following enquiry with City of Montreal |
| MTM8\_Y | DECIMAL | 8,1 | N | Not defined and no response following enquiry with City of Montreal |

Table 2 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 |  |

Table 3 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 |  |

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

Table 4 transactions\_deneigement\_saison\_2018-2019, csv file

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Name** | **Type** | **Length** | **Nullable** | **Comments** |
| TransactionID | INT |  | N |  |
| DateChargement | TIMESTAMP | 19 | N | Date format: 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 |  |

Table 5 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 | Unknown type, unknown length – not used by our system |
| Max Temp | DECIMAL | 3,1 | Y |  |
| Max Temp Flag |  |  | Y | Unknown type, unknown length – not used by our system |
| Min Temp (°C) | DECIMAL | 3,1 | Y |  |
| Min Temp Flag |  |  | Y | Unknown type, unknown length – not used by our system |
| Mean Temp (°C) | DECIMAL | 3,1 | Y |  |
| Mean Temp Flag |  |  | Y | Unknown type, unknown length – not used by our system |
| Heat Deg Days (°C) | DECIMAL | 3,1 | Y |  |
| Heat Deg Days Flag |  |  | Y | Unknown type, unknown length – not used by our system |
| Cool Deg Days (°C) | DECIMAL | 3,1 | Y | Always 0 |
| Cool Deg Days Flag |  |  | Y | Always blank |
| Total Rain (mm) |  |  | Y | Unknown type, unknown length – not used by our system |
| Total Rain Flag |  |  | Y | Unknown type, unknown length – not used by our system |
| Total Snow (cm) |  |  | Y | Unknown type, unknown length – not used by our system |
| 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 | Unknown type, unknown length – not used by our system |
| 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

Table 6 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 |

Table 7 Secteurs

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

Table 8 Arrondissement

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

Table 9 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 |  |

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

Table 10 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 |

Table 11 Weather

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

## Appendix A: Acronyms

Table 12 - Acronyms

| Acronym | Literal Translation |
| --- | --- |
| ACID | Atomicity Consistency Isolation Durability |
| API | Application Programming Interface |
| BASE | Basic Availability Soft-State Eventual Consistency |
| COM | City of Montreal |
| CRD | Conceptual Relationship Data Model |
| csv | Comma Separated Values (file format) |
| DBMS | Database Management System |
| DDL | Data Definition Language |
| EC | Environment Canada |
| EDT | Eastern Daylight Time |
| ERD | Entity Relationship Diagram |
| ERD | Entity Relationship Model |
| ETL | Extraction Transformation Load |
| FK | Foreign Key |
| GB | GigaBytes |
| GPS | Global Positioning System |
| HDD | Hard Drive Disk |
| LDM | Logical Data Model |
| LUKS | Linux Unified Key Setup |
| MB | MegaBytes |
| PK | Primary Key |
| RAM | Random Access Memory |
| RDBMS | Relational DBMS |
| SDK | Software Development Kit |
| SIT-Neige | Système Intelligent de Transport de la Neige |
| SLA | Service Level Agreement |
| SQL | Structured Query Language |
| SSD | Solid State Drive |
| SSL | Secure Sockets Layer |

## Appendix B: DDL

## Appendix C: Decision chart database system

Scofield, Ben (2010-01-14). [10] ["NoSQL - Death to Relational Databases(?)"](http://www.slideshare.net/bscofield/nosql-codemash-2010)

Table 2 – Database options

| **Data model** | **Performance** | **Scalability** | **Flexibility** | **Complexity** | **Functionality** |
| --- | --- | --- | --- | --- | --- |
| Key–value store | high | high | high | none | variable (none) |
| Column-oriented store | high | high | moderate | low | minimal |
| Document-oriented store | high | variable (high) | high | low | variable (low) |
| Graph database | variable | variable | high | high | [graph theory](https://en.wikipedia.org/wiki/Graph_theory" \o "Graph theory) |
| Relational database | variable | variable | low | moderate | [relational algebra](https://en.wikipedia.org/wiki/Relational_algebra" \o "Relational algebra) |

**Key-value stores** are capable of providing much higher performances than RDBMS

**>>** In our case data organization and management is more important than the performance.

**Column family** databases are designed for large volumes of data, read and write performance, and high availability

**>>** We don’t have large volumes and don’t need great performance nor high availability.

**Document databases** have ability to store varying attributes along with large amounts of data

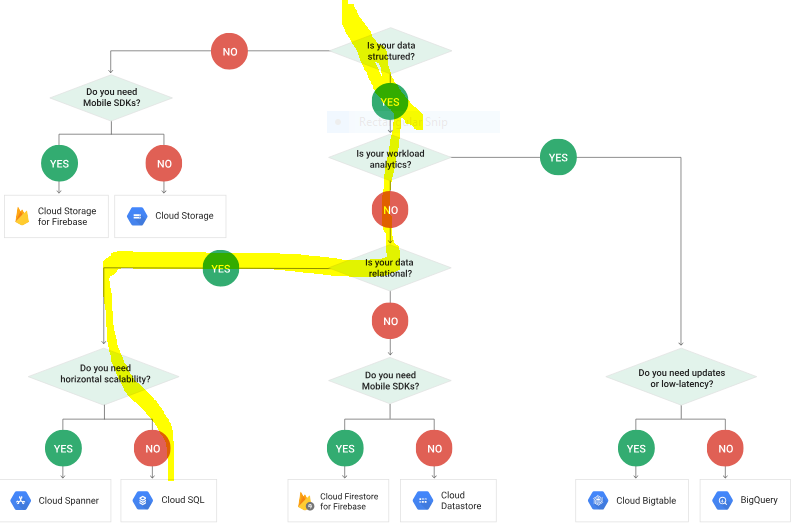
**>>** We don’t have large volumes and the data is static (format doesn’t change much).

**Graph database** is designed to treat the relationships between data as equally important to the data itself. It is intended to hold data without constricting it to a pre-defined model.

**>>** The relationships between the data is static and the pre-defined model does not change.

## Appendix D: Decision tree

[11] <https://cloud.google.com/storage-options/>



## Appendix E: Highlight of milestone meetings

March 7th: Review of Daniel’s feedback on Slack concerning the report. Update of milestones. Review of Digital Ocean and Microsoft Azure’s pricing. Data flow diagram updated on Real Time board. Discussion concerning the difference between the object-relational database and the entity relational database.

March 5th: Debrief following comments from Daniel in class concerning chosen path for Acid and Relation database with SQL. Review of undefined sections. No response from city of Montreal for measurement questions. Data flow diagram updated on Real Time board, and decision on the type of database (columnar, object vs relational).

February 28th: meeting to review Amazon and Google database options (storage, and Acid/Base characteristics), follow-up on next steps No response from city of Montreal for measurement questions. ERD diagram added to the report. Submission of our questions and report to Daniel on Slack.

February 24th: Discuss open data resources available, similar projects, and streets of Montreal surface area. No response received from city of Montreal, following another attempt made using the data request form. Review of locations in Montreal without depots. Start and end dates of season for snow removal contracts are not confirmed by the city of Montreal. No response from city of Montreal for measurement questions.

February 22nd: No response from city of Montreal for measurement questions concerning the volume measure and the start of the season.

February 19th: Review of video communicated by the city of Montreal concerning the snow collection.

February 13th: Discussion concerning license for city of Montreal data and realism in the application of the proposed solution. Review of Système Planif-Neige API document. Discussion concerning the hierarchy of the data and the data dictionary. Data dictionary including data type updated in project document. Use of Tableau to visualize the boroughs were depots and snow collection transactions are being recorded.

February 11th: meeting kick-off and review of Real Time board contents. Sample of database downloaded, and conducted data profiling, in order to access which data to keep. Initial preparation of the data dictionary.

## Appendix F: Decision Chart Storage

Based on pricing and features shown on Capterra [7\*], Digital Ocean [8\*], and Cloudways [9\*].

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Solutions** | | | |
| **Features** | **Zoho** | **Ninox** | **Caspio** | **Digital Ocean** |
| Backup & Recovery | **🗸** | **🗸** | **🗸** | **🗸** |
| Data Migration | **🗸** | **🗸** | **🗸** | **🗸** |
| Data Replication | **🗸** |  | **🗸** | **🗸** |
| Data Security | **🗸** |  | **🗸** | **🗸** |
| Database Conversion | **🗸** |  | **🗸** |  |
| Mobile Access |  | **🗸** |  |  |
| Performance Analysis | **🗸** |  | **🗸** |  |
| Queries | **🗸** | **🗸** | **🗸** | **🗸** |
| Relational | **🗸** | **🗸** | **🗸** | **🗸** |
| Virtualization |  |  | **🗸** |  |
| **Price/Month for 10 users** | $150.00 | $83.30 | $59.00 | $25.00 |

## References

[1\*] http://ville.montreal.qc.ca/snowremoval/operations-delais

[2\*] <https://umq.qc.ca/publication/montreal-deneigement-intelligent/>

[3\*] http://donnees.ville.montreal.qc.ca/dataset/contrats-transaction-deneigement

[1] <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>

[2] <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>

[3\*]

[3] <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>

[4\*] http://ville.montreal.qc.ca/snowremoval/elimination-neige#carte-elimination

[4] <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>

[5\*] <https://www.spiceworks.com/>

[5] <http://climate.weather.gc.ca/historical_data/search_historic_data_e.html>

[6] <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>

[7] <https://www.digitalocean.com/products/managed-databases/>

[7\*]

[8] <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>

[8\*]

[9] <http://donnees.ville.montreal.qc.ca/dataset/contrats-transaction-deneigement>

[9\*] <https://www.cloudways.com/en/pricing.php#digitalocean>

[10] ["NoSQL - Death to Relational Databases(?)"](http://www.slideshare.net/bscofield/nosql-codemash-2010)

[10\*] <https://www.digitalocean.com/community/tutorials/how-to-create-an-encrypted-file-system-on-a-digitalocean-block-storage-volume>

[11] <https://cloud.google.com/storage-options/>