Snowstorm

# Database Design Document

1. Version 0.9
2. 03/21/2019
3. Team: **Ilia Kassianenko, Pawel Kaluski, Fritz Gyger**

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

**CEBD-1250 – Big Data Storage**

Table of Contents

[1. Overview 4](#_Toc4082380)

[1.1 Scope 4](#_Toc4082381)

[1.2 Existing applications 4](#_Toc4082382)

[1.3 Data acquisition 4](#_Toc4082383)

[1.4 Contexts for Analytics 4](#_Toc4082384)

[1.5 Timeframe 4](#_Toc4082385)

[1.6 Base Property for Data 4](#_Toc4082386)

[2. Assumptions/Constraints/Risks 4](#_Toc4082393)

[2.1 Assumptions 4](#_Toc4082394)

[2.1.1 Ville de Montréal 4](#_Toc4082397)

[2.1.2 Environment Canada 4](#_Toc4082400)

[2.2 Constraints 4](#_Toc4082402)

[2.3 Risks 4](#_Toc4082403)

[3. Design Decisions 4](#_Toc4082404)

[3.1 Key Factors Influencing Design 4](#_Toc4082412)

[3.2 Functional Design Decisions 4](#_Toc4082413)

[3.3 Database Management System Decisions 4](#_Toc4082414)

[3.4 Security and Privacy Design Decisions 4](#_Toc4082416)

[3.5 Performance and Maintenance Design Decisions 4](#_Toc4082421)

[3.5.1 Service Level Agreements 4](#_Toc4082422)

[3.5.2 Backup and restoration 4](#_Toc4082423)

[4. Detailed Database Design 4](#_Toc4082431)

[4.1 Roles and Responsibilities 4](#_Toc4082432)

[4.2 Performance Monitoring and Database Efficiency 4](#_Toc4082433)

[4.2.1 Operational Implications 4](#_Toc4082434)

[4.2.2 Data Transfer Requirements 4](#_Toc4082436)

[4.2.3 Data Formats / Data Dictionary 4](#_Toc4082438)

[5. Appendix A: Acronyms 4](#_Toc4082461)

[6. Appendix B: DDL 4](#_Toc4082462)

[7. Appendix C: Decision chart database system 4](#_Toc4082463)

[8. Appendix D: Decision tree 4](#_Toc4082464)

[9. Appendix E: Highlight of milestone meetings 4](#_Toc4082465)

[10. Appendix F: Decision Chart Storage 4](#_Toc4082466)

[11. References 4](#_Toc4082467)

## 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.

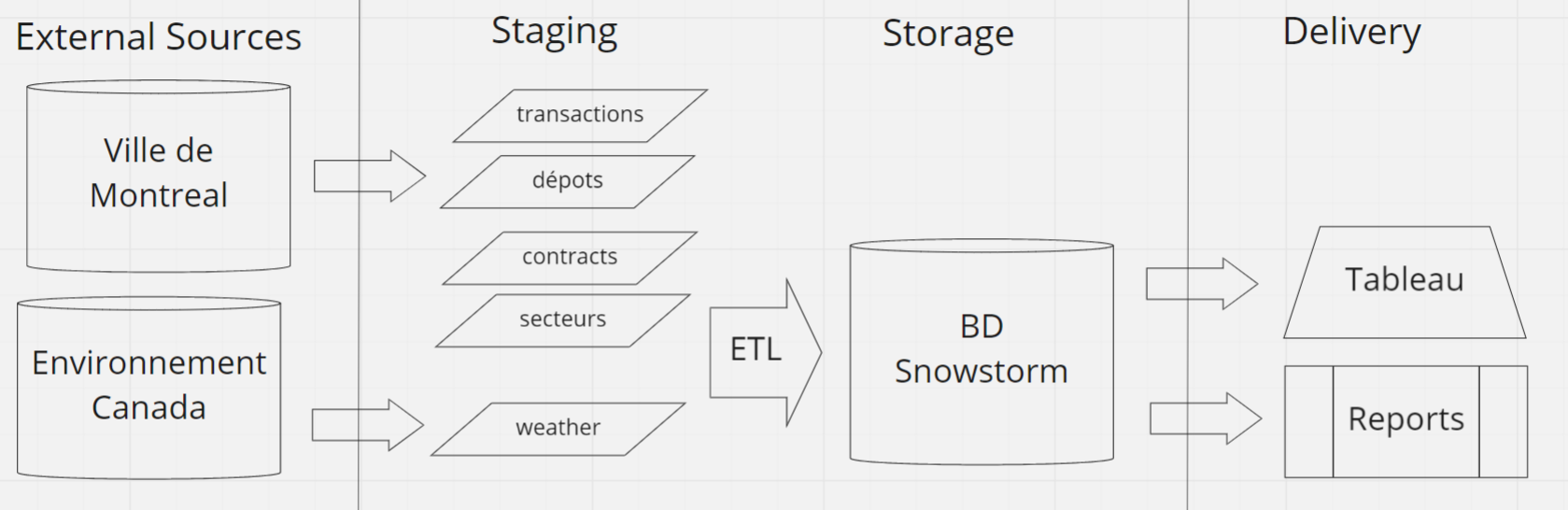


Figure 1 Data Sources

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.

Snow volume, which is unloaded daily, at each depot, is divided by the daily average precipitation depth. The result from this division represents an estimate for the daily surface area, which has received precipitation and which 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

The proposal accounts for a minimum of 10 of history, in order to compare each year relative to the prior and to able observe any changes to a trend.

### 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 new data (CSV files) 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.

## Assumptions/Constraints/Risks

### Assumptions

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. Precipitation is recorded for both snow and rainfall, while only the snow depth is considered. The impact of rainfall on snow, in terms of added weight at the time of snow collection, is not considered.

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

* 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

* 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

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

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 and section 3.5

### Functional Design Decisions

**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 the corresponding rules in [sections 4.2.3.2 and 4.2.3.3](#_Validation_rules:).

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

The proposed solution, Digital Ocean, is considered to offer the needed functionality at the lowest price. Should there be a change of magnitude to the price or any other issue with Digital Ocean, the proposed alternative is Caspio.

information, whichresulted in this assessment, are shown within .

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

Digital Ocean offers PostgreSQL, which is “SQL-compliant [and an] open source object-relational database system”. [7]

Digital Ocean enables our team to “access and manage [our] databases through [its] simplified UI or programmatically through [its] API”. [7] It also handles maintenance operations and updates, and its offers free technical support around the clock. [7]

### Security and Privacy Design Decisions

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 in place for Digital Ocean. [11\*]

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 website.

### Performance and Maintenance Design Decisions

#### Service Level Agreements

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 weekday.
* Normal site operations will occur during weekdays, 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 integration to Spiceworks [5\*].
* Resolution of support tickets will occur the following work day of the week.
* This permits downtime for operations and maintenance, for archiving purposes and for automated ETL processes for the source files.

#### Backup and restoration

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 to have a warm property.

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 propose an initial use of a single node managed database cluster with 1GB of RAM. Digital Ocean offers automatic failback with standby nodes, in case of node failure, in order for data handling to be automatically switched to a standby node. [12\*] A 1 GB RAM PostgreSQL node with a “max\_connections” parameter set to 22 connections, where 3 connections are reserves for maintenance. [12]

Weekly automatic updates for the database engine and the operating system will occur during nighttime, in order to keep the service stable.

## Detailed Database Design

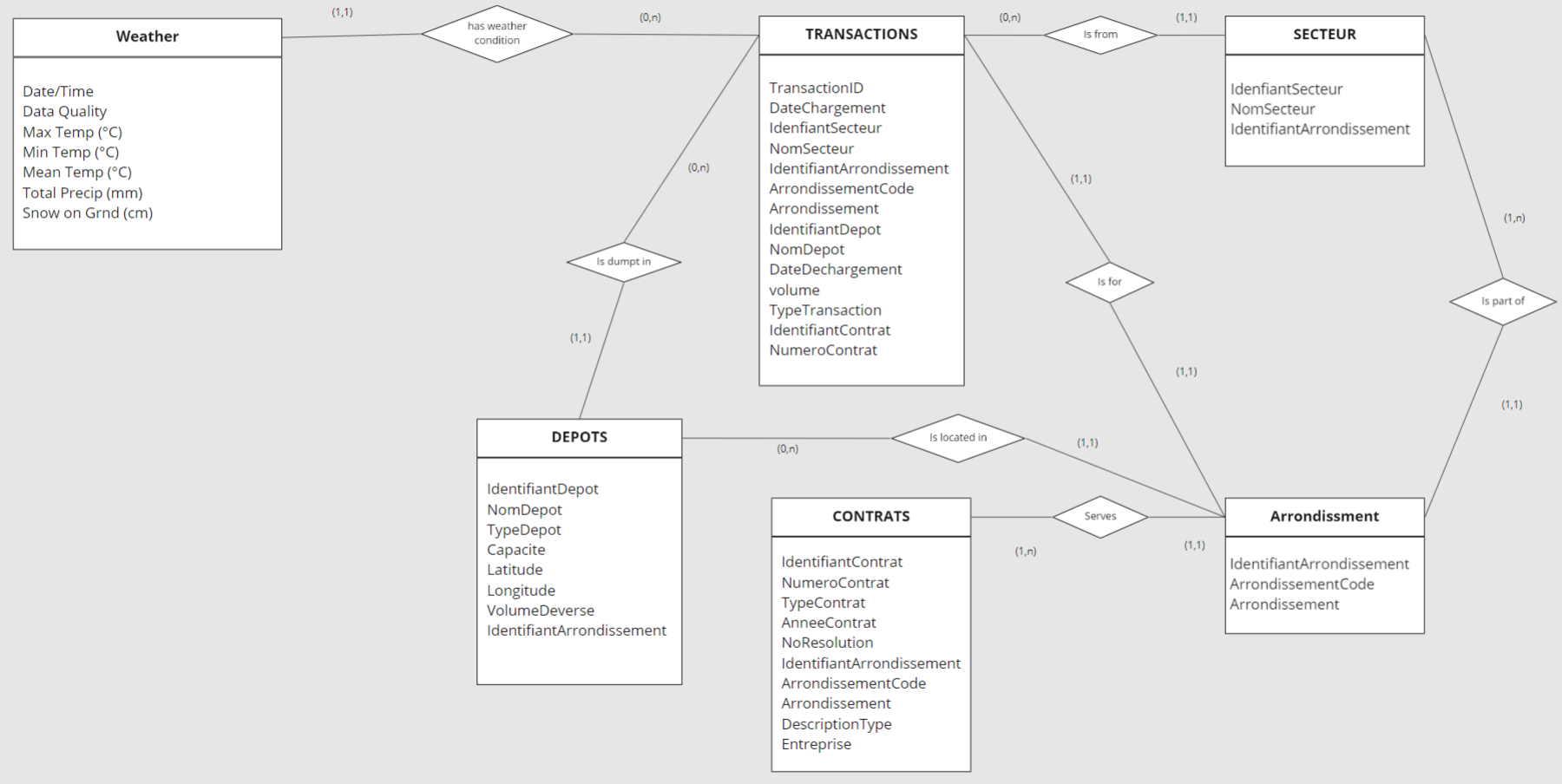


Figure 2 Conceptual Data Model (CDM) [8]

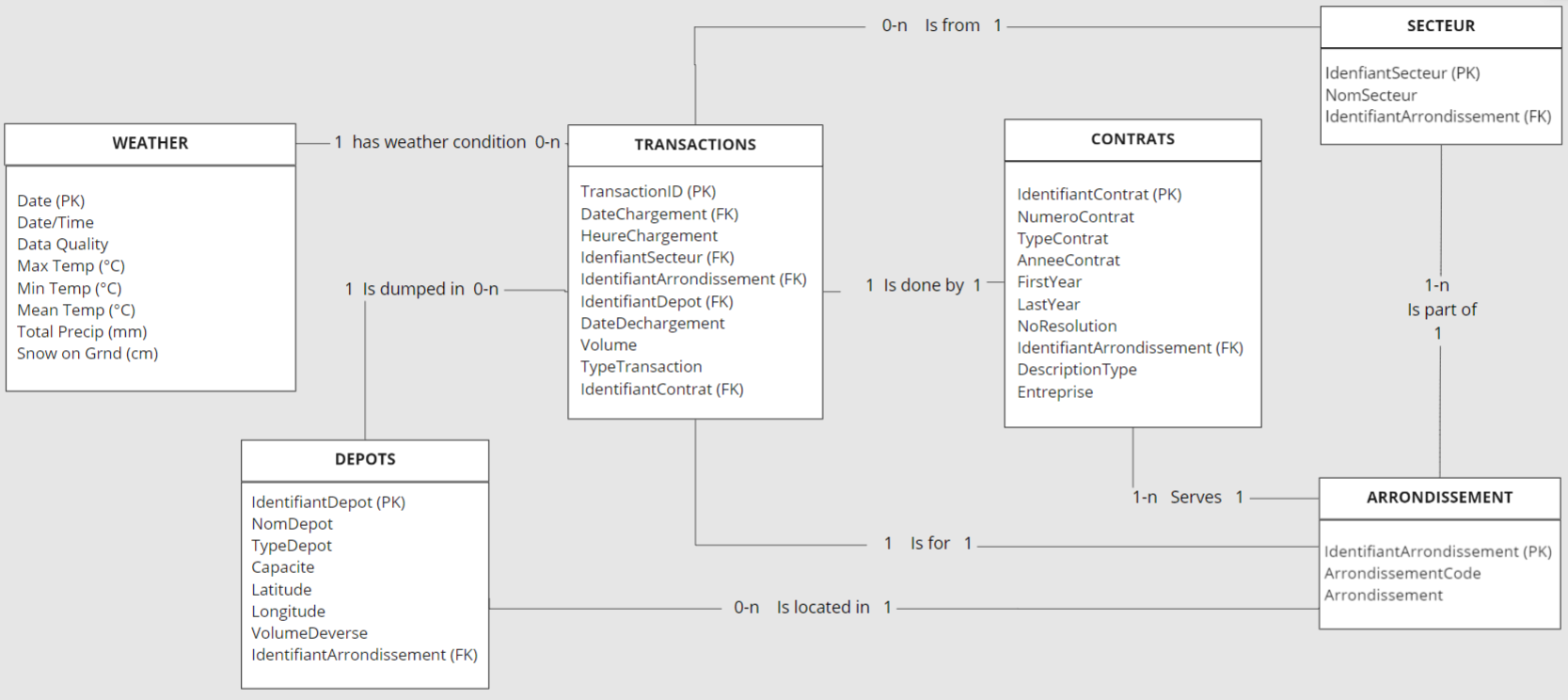


Figure 3 Logical Data Model (LDM) and LDM Entity Relationship Diagram (ERD). [8]

* Physical Data Model (PDM) with a description of the DBMS schemas, sub-schemas, records, sets, tables (DDL) :

See Appendix B (DDLs)

* Expected volume :
  + Initial: approximately 35 MB
  + Weekly growth: about 10 MB for transaction and weather data.
* Update frequency :
  + Weekly Mondays after 1 AM EST: the COM updates the files every Monday at 1am EST
  + Weather Mondays after 1 AM EST

### Roles and Responsibilities

* Creation and maintenance of tables, etc.: All three team members
* Database administration (backups, etc.) : Digital Ocean, as a third-party
* Security administration: All three team members, and Digital Ocean, as a third-party
* Data Acquisition & cleansing (ETL): All three team member, with a continuing education for a member willing to become a Data Engineer
* Reports & analytics: All three team member, with a continuing education for a member willing to become a Data Scientist

### Performance Monitoring and Database Efficiency

#### Operational Implications

* Transactions & Weather:
  + COM data: Weekly Mondays after 1 AM EST, as the COM updates their files at this time.
  + Weather data: We propose to extract the data also on Mondays after 1 AM EST, in order to synchronize the extractions and facilitate the ETL processes.

#### Data Transfer Requirements

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

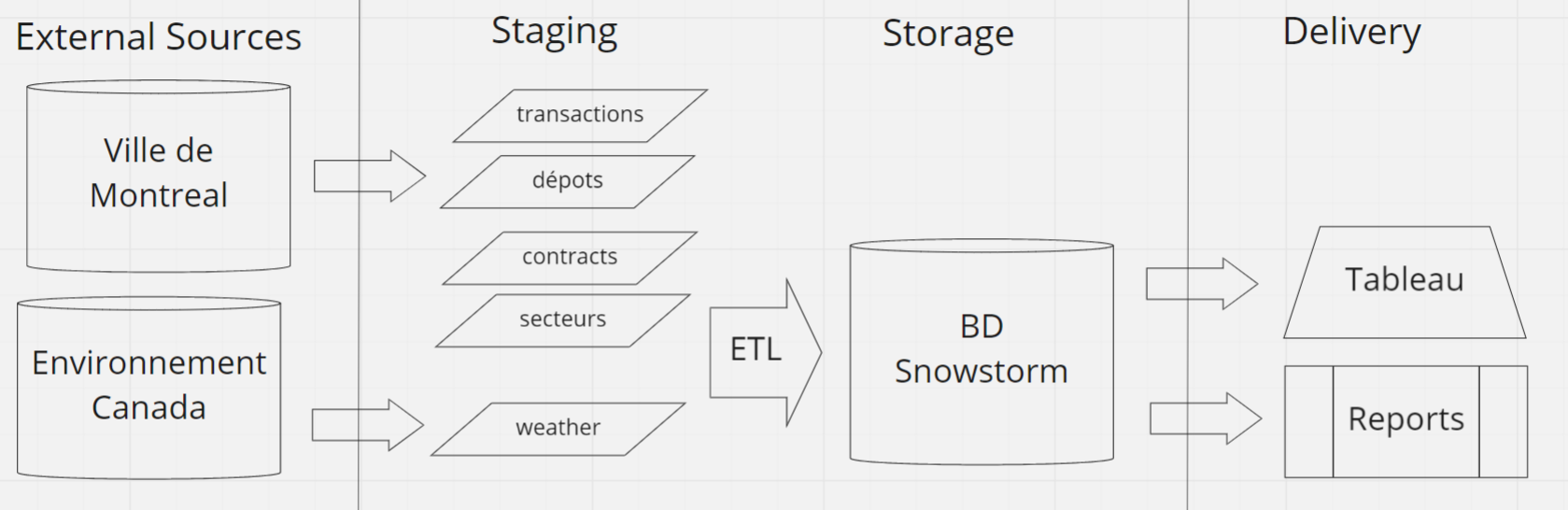


Figure 4 Data Sources

* Extract source files in UTF-8 format in order to have resolve string conversion errors for French accents

(example: “Saint-Léonard” instead of “Saint-LÃ©onard”)

* Volume:
  + Initial: about 35 MB
  + Weekly growth: about 10 MB for transaction and weather data.
* Data content and format (Refer to section [4.2.3 Data Formats](#_Data_Formats_/))

#### Data Formats / Data Dictionary

Based on the COM dictionary [9], the following information is available.

#### 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 | Date format: 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 (example: “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 | Capacity of “0” for a 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  Date format: 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  Date format: YYYY-MM-DD |
| Max Temp | DECIMAL | 3,1 | Y | Celsius, degrees |
| Min Temp | DECIMAL | 3,1 | Y | Celsius, degrees |
| Mean Temp | DECIMAL | 3,1 | Y | Celsius, degrees |
| 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 |
| URL | Uniform Resource Locator |

## Appendix B: DDL

## Appendix C: Decision chart database system

Table 2 – Database options [10]

| **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 are more important than the performance.

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

**>>** We do not have large volumes, do not need great performance, and do not need high availability.

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

**>>** We do not have large volumes, and the data is static, as the format does not 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

Storage options decision tree [11]

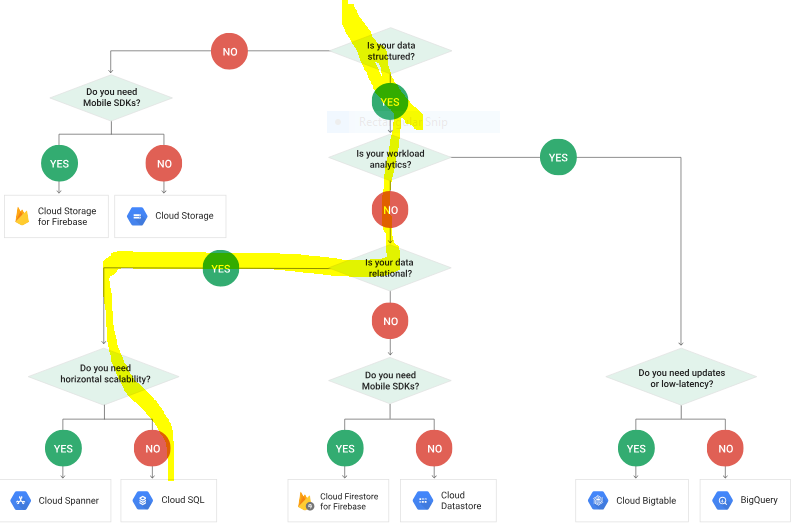


Figure 5 Decision Tree

## Appendix E: Highlight of milestone meetings

March 20th: Debrief on editing changes for the report and update on references.

March 13th: Discuss how the presentation will be split between team members, review of alternatives to digital ocean (on Product Hunt and Capterra); risks related individual start-up and venture backed companies as well as systemic risks (e.g. Nasdaq crash fo 2001) where funding is interrupted for the majority of software providers. The chosen company may increases prices to remain solvent, or the company may declare insolvency and announce an operational closure within two weeks. Theses risks are considered, since there is no credit rate or other means to gage the longterm reliability of the company. In case of solvency, the operational stability of the company servers is assumed to be acceptable.

March 11th: Transition to Python as an ETL language rather than Tableau public, which can be hidden behind a paywall for public users. For further analysis, python can also help prepare the data entering into Hadoop. Server access to URL [ftp.tor.ec.gc.ca](http://ftp.tor.ec.gc.ca) is not permitted as a guest; unknown process for becoming a registered users, in order to access historical records other than in XML or CSV format.

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\*].

Table 13 Decision Chart

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **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 ($CAD)** | $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)Scofield, Ben (2010-01-14).

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

[11\*] https://www.digitalocean.com

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

[12] <https://www.digitalocean.com/community/tutorials/managed-databases-connection-pools-and-postgresql-benchmarking-using-pgbench>

[12\*] https://www.digitalocean.com/docs/databases/how-to/clusters/add-standby-nodes/