[Logo

Description automatically generated](https://twitter.com/ad_hull)[](https://www.linkedin.com/in/alfredhull/)[](https://github.com/ahull002/SP_data_engineering_open-ended-capstone)

**Springboard**

**Ste 1100 22 Battery Street**

**San Francisco, CA 94111-5525**

*Prepared By: Alfred D. Hull*

Data Engineering

Open-Ended Capstone

December 2020



2020

~ This Page is intentionally left blank ~

**Contents**

[**Artifact 1: Project Ideation** 0](#_Toc58436122)

[**Artifact 2: Project Proposal** 1](#_Toc58436123)

[**What is this?** 1](#_Toc58436124)

[**How to use this?** 1](#_Toc58436125)

[**How to test & develop this?** 1](#_Toc58436126)

[**Constraints, Limitations, and Assumptions** 2](#_Toc58436127)

[**Project Setup** 3](#_Toc58436128)

[**Artifact 3: Data Collection & Design Method** 5](#_Toc58436129)

[**Entity Relationship Diagram:** 5](#_Toc58436130)

[**Conceptual Diagram** 5](#_Toc58436131)

[**Functional Diagram** 7](#_Toc58436132)

[**Physical Diagram** 7](#_Toc58436133)

[**Database Design Selection & Assumptions** 8](#_Toc58436134)

[**Deciding Factors** 8](#_Toc58436135)

[**References:** 0](#_Toc58436136)

[**Appendix:** 1](#_Toc58436137)

~ This Page is intentionally left blank ~

# **Artifact 1: Project Ideation**

## **Artifact deliverables:**

* Choose a dataset that is at least 100s of GigaBytes in size
* Provide a quick write-up of the concept for the project

### **Chosen Data Set:**

Data: New York Taxi & Limousine Commission data set. Link: <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

### **Concept Summary:**

This project aims to process the New York Taxi & Limousine data to be analyzed and pipelined to a cloud database. This tool extracts trip data from taxi cabs around the city of New York. It processes it using sophisticated algorithms to make sense of the data and provide analysts with easy to understand capabilities to build educational data products in the cloud. An example of a data product is a multi-dimensional data cube that enables graphical maps to highlight all the places where trips are most common and expected on any particular day of the week. For instance, one would be able to analyze where customers are dropping off and picking up patrons throughout the day at certain times of the day, along with the histories of these transactions done over recent months. Staging tables in this software application will enable the operator to process data in many different ways through connecting business intelligence or data science software. If analysts want to know how long a driver has been driving in any particular taxi zone, you can look up the driver's record in the software's database and see how they have performed. You can also lookup a specific taxi zone and see whether it is experiencing customer satisfaction problems. Once you know what is going on in your local taxi area in the taxi zones, you can then plan new routes or re-arrange cabs to improve efficiency and profitability.

Taken as a whole, the complete detailed trip-level data from a taxi ride is much more than a vast array of pickup and drop off points: it is a window into the city of New York, detailing every last detail of the journey. Data-engines will do the job of extracting this information and presenting it in a useful form. Data-engines are software applications that use large-scale data analysis tools and data visualization capabilities to allow people and companies like you and me access to information that has been processed and stored by sophisticated algorithms.

We believe building this tool will open up another avenue for decision support capabilities.

~ This Page is intentionally left blank ~

# **Artifact 2: Project Proposal**

*This phase defines the opportunity this project intends to address, clarifies expected outcomes, and highlights the necessary resources required for optimal success.*

## **Artifact deliverables:**

* Explanation of the project
* How to use it
* How to test & develop it
* Constraints, Limitations, and Assumptions
* Project setup

### **What is this?**

This document includes the write-up for a Data Engineering project capstone at Springboard that covers engineering an en- to-end data pipeline using the NYC Taxi and Limousine Commission Data. The capstone highlights a series of data engineering steps required to build data flows. The pipeline computationally ingests data at the pipeline's beginning; there are many steps in which each step in the pipeline delivers an output-input to the next step. The process highlighted here continues until the channel is complete. In some cases, independent efforts run in parallel.

This data pipeline consists of three key elements: a source, a processing step or steps, and a destination (a sink). The data pipeline enables data to flow from an application to a data warehouse (MS Azure) for future analytics and report generations.

### **How to use this?**

This data pipeline is architected as a batch-based flow. In this architecture scenario, we have built an Extraction Transformation Load (ETL) flow from the New York Taxi & Limousine Commission data sets to push this data to a data warehouse and an analytics database.

### **How to test & develop this?**

This data pipeline is architected as a batch-based flow. In this architecture scenario, we have built an Extraction Transformation Load (ETL) flow from the New York Taxi & Limousine Commission data sets to push this data to a data warehouse and an analytics database.

### **Constraints, Limitations, and Assumptions**

**Question 1. Data sources you are considering for your open-ended capstone**

***Data****: New York Taxi & Limousine Commission data set.*

***Link****: https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page*

**Question 2. Volume available of data for each source (historic and delta)**

The historical load comprises the below-archived file sets for each month of the year since 2009, totaling around 288 gigabytes. The complementary delta loads occur each month and increment the data set by at least 2 gigabytes in total.

***Delta Load (2 gigabytes per month).***

*Yellow Taxi Trip Records (CSV): file size – 579,698 KB*

*Green Taxi Trip Records (CSV): file size – 39,744 KB*

*For-Hire Vehicle Trip Records (CSV) – 108,186 KB*

*High Volume For-Hire Vehicle Trip Records (CSV) – 1,273,342 KB*

***Historical Load (288 gigabytes).***

*Yellow Taxi Trip Records (CSV): file size – 83,476,512 KB*

*Green Taxi Trip Records (CSV): file size – 5,723,136 KB*

*For-Hire Vehicle Trip Records (CSV) – 15,578,784 KB*

*High Volume For-Hire Vehicle Trip Records (CSV) – 183,361,248 KB*

**Question 3. The volume you will be using for your capstone**

As of December 2020, I plan on utilizing all 288 gigabytes of the Historical Load, and I will delta load the new storage location by the number of months from December to the month when the capstone is due.

**Question 4. Why do you think this is a good data source to be used for the capstone**

I believe this is a good data source as the data is extensive and has large delta loads that are still active and will help people analyze the end product versus spending time wrangling the data itself.

**Project Pros & Cons:**

**Pros**

* As of December 2020, the NYC Taxi and Limousine Commission's actively maintains the file sets and protect them under the Freedom Of Information Law (FOIL).
* The US NYC Government is transparent in sharing this information with the general public.

**Cons**

* There is no API built for the data sets, and several folders partition the data sets out by year and month.
* The file format is in .csv, and there may be risks associated with file corruption.

We will use this project management model to ensure that the client can save time and money throughout the process and have the flexibility to make changes anytime during the development process.

### **Project Setup**

The NYC Taxi & Limousine Commission, Data Engineering Project (NTLCDEP), will use the Agile Methodology as the project management process. The methodology involves breaking down each project into prioritized requirements and delivering each piece within an iterative cycle. Each iteration is evaluated and assessed (sprint retrospective) by the development team and the client (Springboard Mentor). The insights gained from the retrospective determines development next steps. We set clear goals in each iteration meeting during the retrospective, such as; expected changes, time estimates, priorities, and budgets.

The agile method gives high priority to client participation at the project's initiation and throughout its lifecycle. The intention is to keep the client involved at every step and to incrementally create a product that they are satisfied with at the end. This approach is essential in saving the client money and time because the client tests and approves the work at each development step, quickly making changes if needed.

**Project Setup Goals:**

* Clarifying needs and outcomes and connecting them to small-behaviors and straightforward
* Breaking down broad overwhelming goals into smaller objectives
* Troubleshooting what is not working well
* Learning from what is working well
* Finding ways to get started easier
* Diminishing the risk of harmful unintentional outcomes
* Illuminating the standards for success and expanding the definition of success to include a more comprehensive range of acceptability when learning.
* Examining your progress in a way that encourages continued progress and emphasizes recovering from setbacks over fearing and avoiding them (self-coaching and self-leading.)

~ This Page is intentionally left blank ~

# **Artifact 3: Data Collection & Design Method**

The design criteria phase will help lay the theoretical engineering solution and a foundational way-forward to establish the data pipeline. This phase is critical as it walks through the conceptual diagram, functional diagram, and physical diagram. This approach is essential so that engineers understand the scope of the build-out and why.

## **Artifact deliverables:**

* Data Collection Flow
* Database Design Selection & Assumptions

### **Database Design Selection & Assumptions**

#### **Deciding Factors**

The engineering team considered two approaches to the database design selection: Inmon and Kimball.

**Inmon**. The Inmon approach considers beginning with a corporate data model. This data model identifies the key subject areas, and most importantly, the fundamental entities, attributes, and connections the business operates within production.

**Kimball**. The Kimball approach considers the critical business processes and the essential business questions needing to be answered by the data warehouse. Additionally, key-sources (operational systems, for instance) generating data for the warehouse are analyzed and documented appropriately. Extraction Transformation and Loading (ETL) software brings data from various sources and load it to a staging area. Data is then parsed and loaded into a dimensional model. The critical difference between Inmon and Kimball's model for the data warehousing technique is that the dimensional model does not require normalization (for instance, the third normalization form (3NF)). The underlying concept of dimensional modeling is the star schema.

1. **Reporting Requirements** – The reporting requirements are not strategic and enterprise-wide, and no integrated reporting is needed. The reporting requirements are tactical in nature and business process/team-oriented.
2. **Project Urgency** – The organization does not have enough time to wait for the first delivery of a robust data warehouse (4 to 9 months). The team has a limited timeline for engineering the data warehouse, standing it up, and running it (2 to 3 months). Historical expertise has shown that the Kimball approach is best in this case (Breslin, 2004).
3. **Future Staffing Plan** – The charter for this project can not afford to have a large engineering team of specialists to maintain the data warehouse. The plan for the group is to be lean and agile.
4. **Frequency of Changes** – We do not expect the reporting requirements to change rapidly, and the source systems are not known to be volatile. Our team will utilize the Kimball approach as the requirements and source systems are relatively stable.
5. **Organization Culture** – While the data warehouse's sponsors and the managers understand the data warehouse's value proposition, this effort is simply a project. It is not seeking long-lasting value from the data warehouse investment. The engineering team is merely prototyping a data engineering pipeline utilizing real-world data as an example to apply methods in this project to other use-cases.

### **Data Collection & Flow**

Graphical user interface

Description automatically generated

**Figure. DB Collection & Flow**

Revise this architecture because you can transform big data in memory. Switch this to Extract Load Transform (ELT) on a intake data mart that then pushes it into the particular schema.

# **Artifact 4: Exploratory Data Analysis**

## **Artifact deliverables:**

* Dimensional Design
* Entity Relationship (ER) Diagrams
* Conceptual
* Functional
* Physical
* Exploratory Data Analysis
* Stored in the docs folder: eda\_fhv\_tripdata\_output
* Stored in the docs folder: eda\_fhfhv\_tripdata\_output
* Stored in the docs folder: eda\_green\_tripdata\_output
* Stored in the docs folder: eda\_yellow\_tripdata\_output

### **Dimensional Design**

#### **BUS Diagram:**

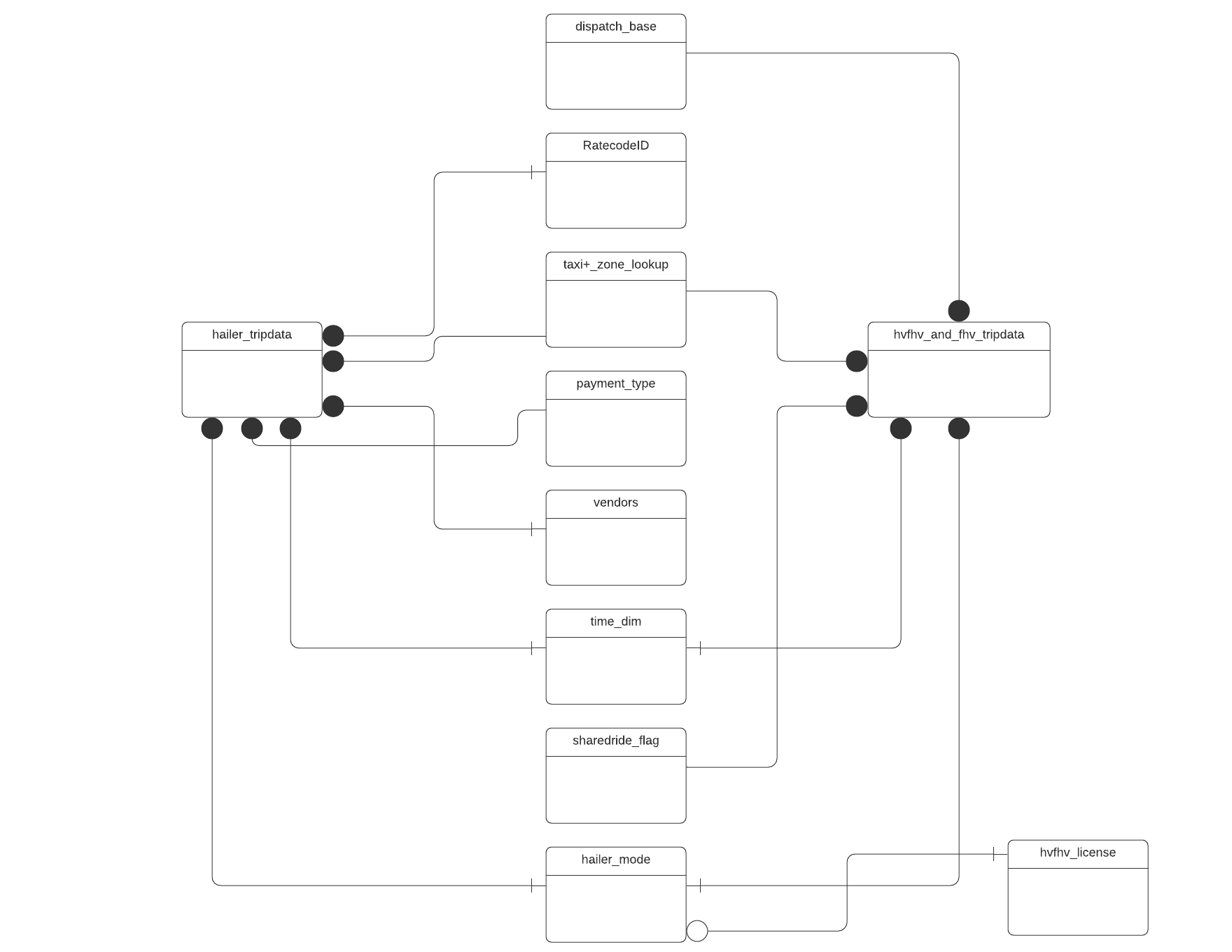
The below BUS Matrix highlight the identified business processes



### **Entity Relationship (ER) Diagram**

#### **Conceptual (ER) Diagram:**

The below first-draft concept diagram allows us to understand the extraction process from the NYC Tax & Limousine data sets. This process is a critical step in understanding how to engineer the ETL pipeline.



**Figure. DB Conceptual Diagram**

#### **Functional (ER) Diagram:**

The below function diagram highlights where tables point within the database schema. This process is a critical step in understanding the storage and movement of the data.

Graphical user interface

Description automatically generated

**Figure. DB Functional Diagram**

#### **Physical ER Diagram:**

The below physical diagram highlights how the database schema will be architected and scripted. Normalization 3F

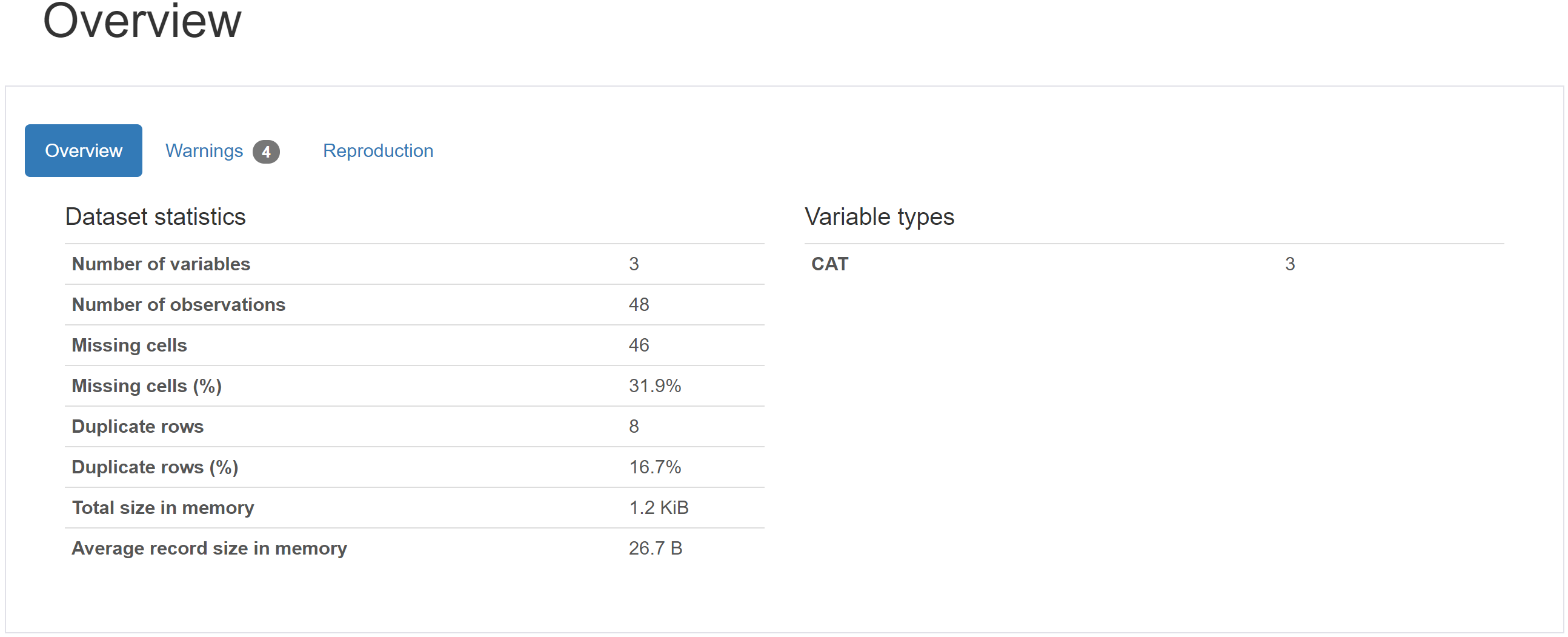
Graphical user interface

Description automatically generated

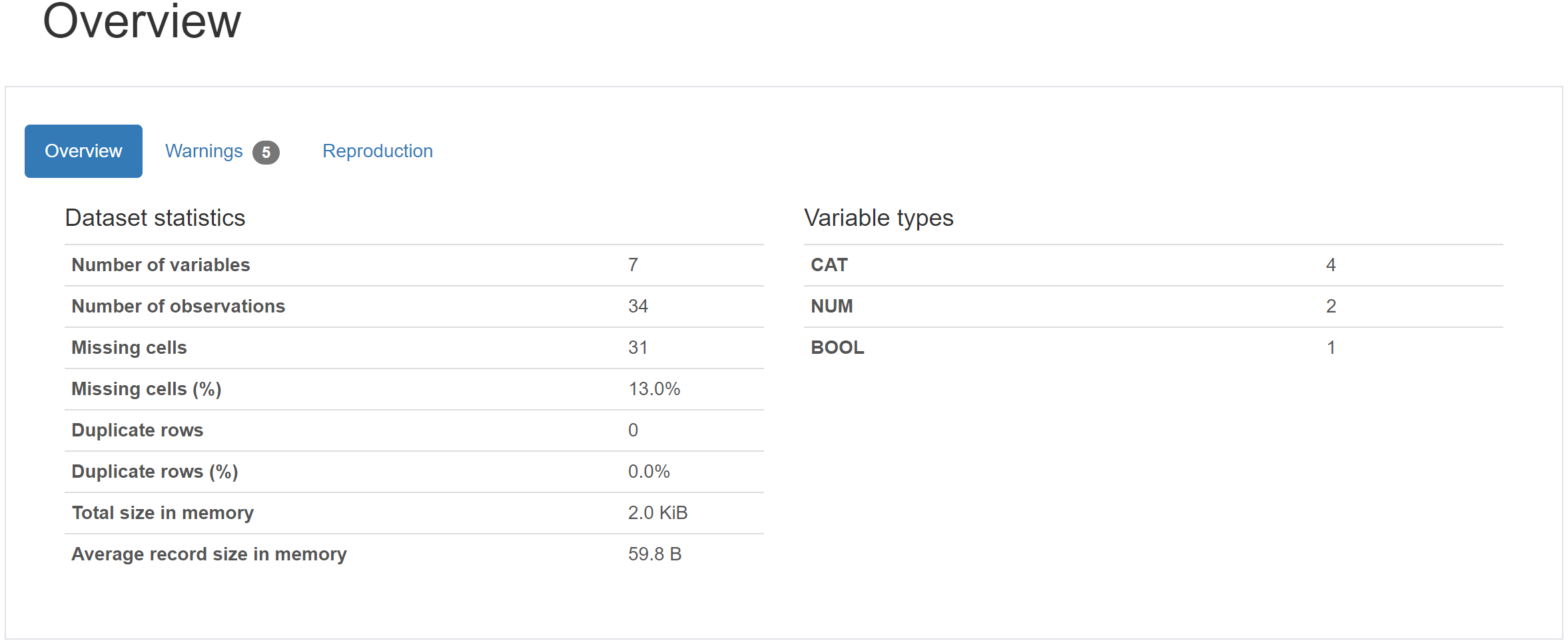
**Figure. DB Physical Diagram**

### **Exploratory Data Analysis**

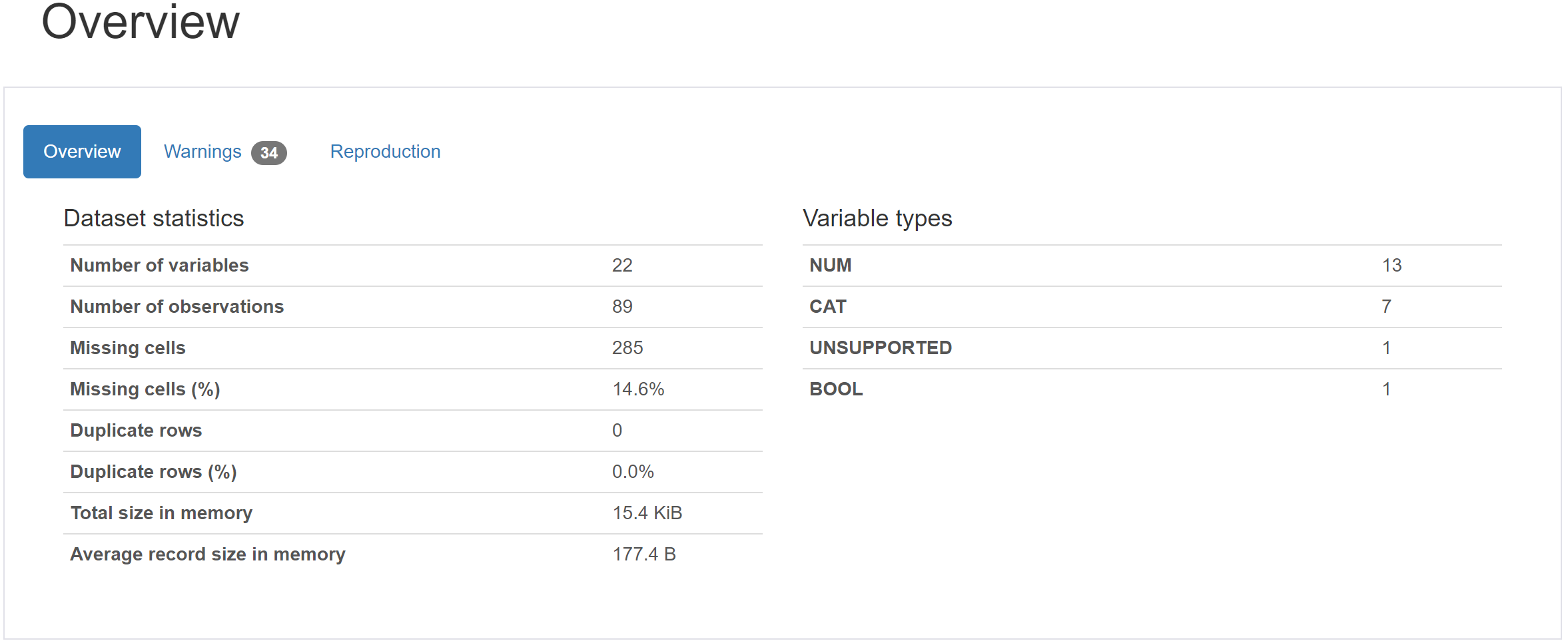
#### **eda\_fhv\_tripdata\_output:**



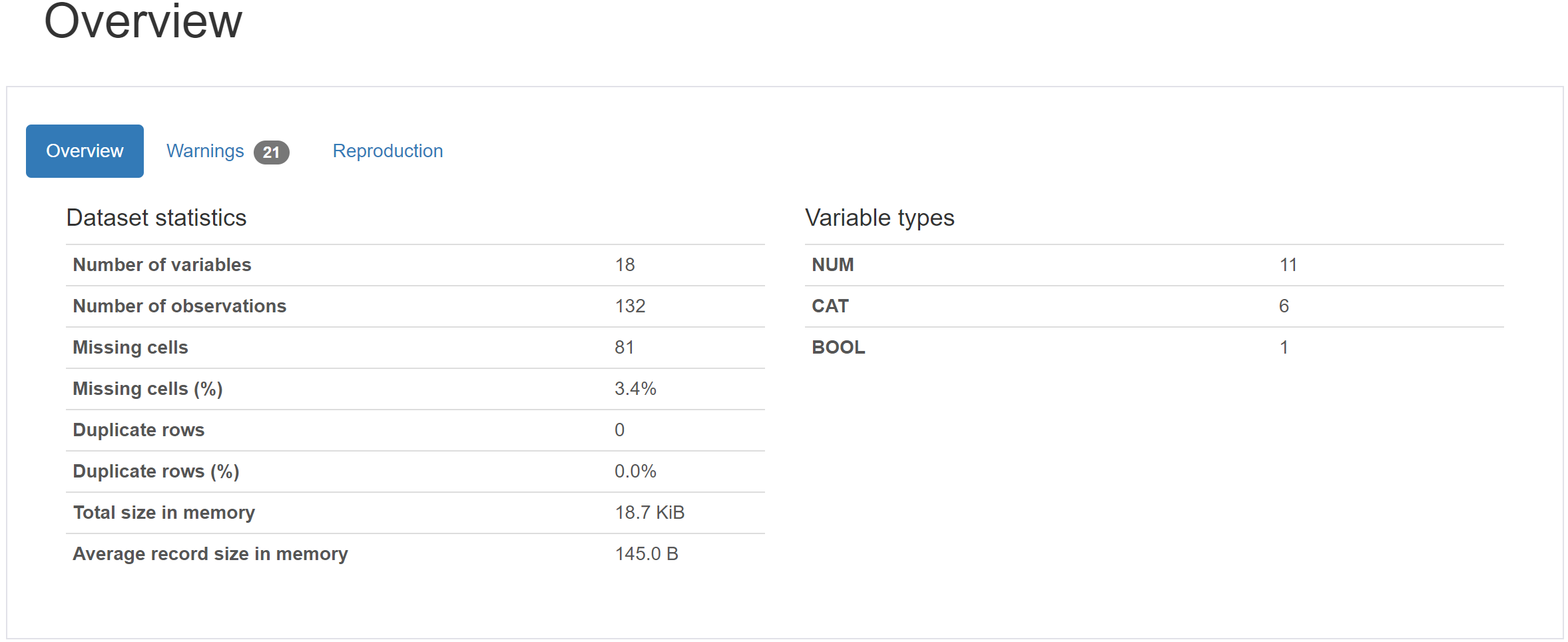
#### **eda\_fhfhv\_tripdata\_output:**



#### **eda\_green\_tripdata\_output:**



#### **eda\_yellow\_tripdata\_output:**



# **Artifact 5: Prototype Pipeline**

## **Artifact deliverables:**

### **Cleaning/Transformation Choices**

Choices I had to make about any cleaning/transformation of the data in my prototype

* TBD
* TBD
* TBD

### **Automation Choices**

Choices I made about the automation of my data pipeline that impact its performance or reliability

* TBD
* TBD
* TBD

# **References:**

Breslin, Mary. 2004. "Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models" Business Intelligence Journal, Winter 2004. Accessed May 22, 2016.

Inmon, W. H. Building the Data Warehouse, Fourth Edition. John Wiley & Sons., 2005.

Marakas, George M. Modern Data Warehousing, Mining, And Visualization. Prentice-Hall, 2003.

Inmon, W. H., 2010. "A TALE OF TWO ARCHITECTURES" InmonCif.com. Accessed May 23, 2016.

http://www.inmoncif.com/products/A%20TALE%20OF%20TWO%20ARCHITECTURES.pdf

Kimball, Ralph, and Margy Ross. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, Third Edition. John Wiley & Sons. 2013. Books24x7.

Stanford. 2003. "Data Warehousing Concepts" Stanford.edu. Accessed May 26, 2016.https://web.stanford.edu/dept/itss/docs/oracle/10g/server.101/b10736/concept.htm#i1006297

Zentut. 2016. "Ralph Kimball Data Warehouse Architecture" Zentut.com. Accessed May 25, 2016. http://www.zentut.com/data-warehouse/ralph-kimball-data-warehouse-architecture/

Owocki, K. (2015, May 9). Data Warehousing #5: Dimensional Design Process. The World As Perpetual Beta. https://owocki.com/data-warehousing-5-dimensional-design-process/

# **Appendix:**

**Open-ended Capstone, Step One**

This subunit will kick off your Open-Ended Capstone! This capstone project will be one of your crucial portfolio pieces. You will create an end-to-end data engineering project over ten steps, from ideation to sourcing data to creating pipelines to hosting it in the cloud. To start, you'll brainstorm a few project ideas to discuss with your mentor.

**Open-ended Capstone, Step Two**

This step of your Open-ended Capstone is focused on creating a proposal for your project. When approaching a data engineering project, it's essential to have a plan of action: to know where you'll get your data from, which tools you plan on using, the problem you are solving, and the project's goals. Some details of the proposal are bound to change as you go through the rest of the steps of the capstone, but the request will be a solid foundation to work from.

**Open-ended Capstone, Step Three**

This step of your Open-ended Capstone will focus on designing your data collection method and then gathering the data you need to begin working on your project. Data collection is a vital part of a data engineer's work!

**Open-ended Capstone, Step four**

Now that you've obtained the dataset, you need to clean, enrich, and transform it. Your dataset

needs to be made ready for consumption by your clients or other members of the data team.

In this step, you will explore your data to understand how it is distributed. Use the lessons from

this exploration to decide how you want to clean and split your data for efficient storage and

querying. You must also create a data model that reduces data redundancy and is also efficient

in terms of storage and querying.

**Open-ended Capstone, Step five**

This step of your Open-ended Capstone will focus on prototyping your data pipeline! This is one of the most important tasks that data engineers do in their daily work. You’ll prototype the pipeline first, and then deploy it to the cloud in later phases. For this step, please use a volume of data that your local computer can handle (2-3 GB). You’ll likely need to slice the dataset in order to select the appropriate volume.