



Maji Ndogo: From analysis to action

Starting the journey

Maji Ndogo: From analysis to action

Welcome to Maji Ndogo

The story we'll step into is not unique to Maji Ndogo; it **mirrors real-world challenges** faced in many places across the globe.

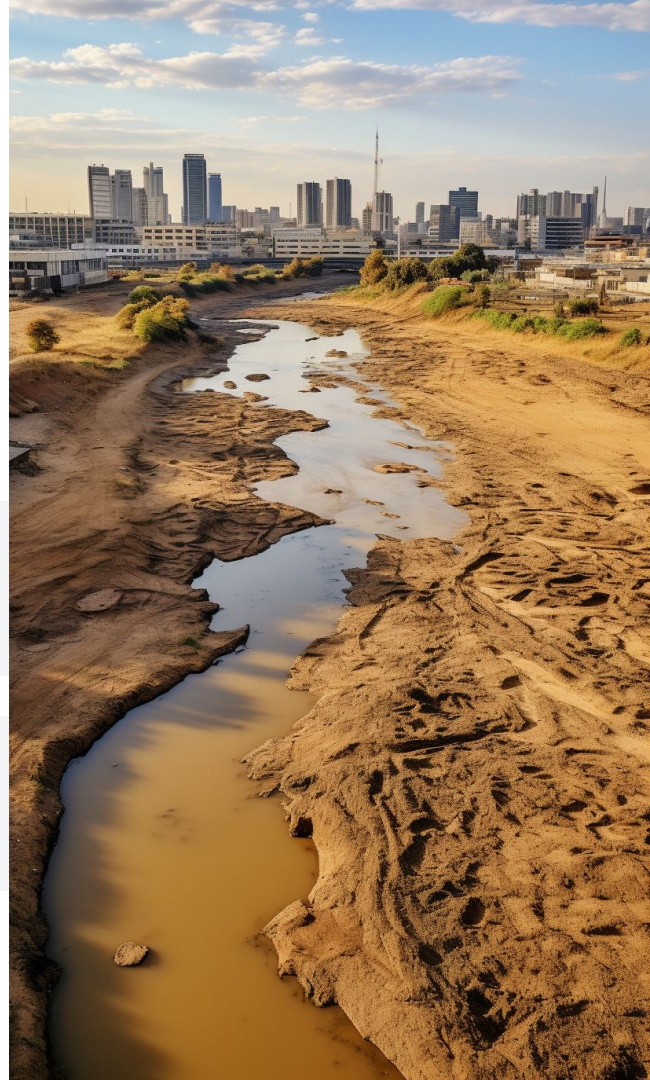
Our mission is to help rejuvenate the drying *Mto wa Matumaini* – The River of Hope – using a data-driven approach.



Completing this mission won't just make you adept at SQL; it will **empower** you to tackle complex challenges, equipping you with highly desirable skills.



Large parts of this project were generated using AI. All characters and places are fictional, but purposely designed.



What will we be doing?

Part 1: Beginning our data-driven journey in Maji Ndogo



Explore a realistic **database with SQL**.



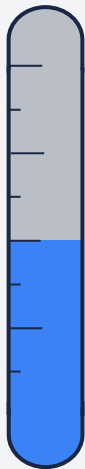
Use SQL to **clean** and **explore** a database with 60,000 unique records.



Become immersed in a story-driven simulation of real data **projects**, **data analysis**, and **good leadership**.

What will we be doing?

Part 2: Clustering data to unveil Maji Ndogo's water crisis



Gear up for a **data analysis** of Maji Ndogo's water scenario.



Harness the power of SQL functions, including intricate window functions, to draw insights from the data.



Aggregate data to unravel the scale of the problem, and start to form some actionable insights.

What will we be doing?

Part 3: Weaving the data threads of Maji Ndogo's narrative



Deal with some of the realities faced in many countries.



Draw from different data sources to deepen the analysis into Maji Ndogo's crisis.



Use advanced SQL tools to assemble the pieces of an audit together.

What will we be doing?

Part 4: Charting the course for Maji Ndogo's water future



Combine SQL tools to finalise our analysis.



Assemble our final analysis results, and report our findings.



Look to the future as we derive actionable goals, and shape the data to achieve them.



Assessments

The integrated project has **four assessment opportunities**. These will be in the format of MCQs and are based on the activities done in that week's part.

Part 1: 10 questions

Create, modify, and explain basic queries related to the water survey database.

Part 2: 10 questions

Use functions to clean, aggregate, and analyse data from the water survey database.

Part 3: 10 questions

Retrieve data from multiple tables in the water survey database.

Part 4: 10 questions

Use functions, filters, and advanced SQL tools to analyse real datasets.

Project instructions format

The format of the slides that **guide us through the project each week** mimics a chat-like interface, similar to Google Chat.

The screenshot shows a chat window with a contact named 'Chidi Kunto' who is 'Online'. The chat history includes a message from Chidi Kunto at 11:05 AM stating they chose two source IDs: 'AkRu85234224' and 'HaZa21742224', and asking to check records for those IDs. A response at 11:09 AM provides a table of data. A follow-up message at 11:11 AM mentions adding more sources and checking queue times.

Introduction
Setting the stage for our data exploration journey.

1. Get to know our data
Exploring the foundational tables and their structure.

2. Dive into sources
Understanding different sources with SELECT.

3. Unpack the visits
Discovering the visit patterns.

4. Water source quality
Understanding water quality.

5. Pollution issues
Correcting pollution data with LIKE and string operations.

I chose these two:
AkRu85234224
HaZa21742224
Ok, so now back to the water_source table. Let's check the records for those source_ids. You can probably remember there is a cool and a "not so cool" way to do it.

This is what I get.

source_id	type_of_water_source	number_of_people_served
AkKi08081224	shared_tap	3398
AkLu01628224	bio_dirty_well	210
AkRu85234224	tap_in_home_broken	496
HaRu19601224	shared_tap	3322
HaZa21742224	poi_dirty_well	388
SoRu36096224	shared_tap	3786
SoRu37635224	shared_tap	3920
SoRu38776224	shared_tap	3180

I added a couple of others... Sorry! Well, if you check them you will see which sources have people queueing. The field surveysors also let us know that they measured sources that had queues a few times to see if the queue time changed.

Project instructions format

The format of the slides that **guide us through the project each week** mimics a chat-like interface, similar to Google Chat.

Chidi Kunto is our best virtual data analyst who will **help to break down tasks** from President Naledi into more technical data questions.

Chidi represents a role model for an analyst. He is a good leader, passionate about the work he does, and critically thinks about everything.

Introduction
Setting the stage for our data exploration journey.

1 Get to know our data
Exploring the foundational tables and their structure.

2 Dive into sources
Understanding different sources with SELECT.

3 Unpack the visits
Discovering the visit patterns.

4 Water source quality
Understanding water quality.

5 Pollution issues
Correcting pollution data with LIKE and string operations.

Chidi Kunto
Online

I chose these two:
AkRu85234224
HaZa21742224
Ok, so now back to the water_source table. Let's check the records for those source_ids. You can probably remember there is a cool and a "not so cool" way to do it.

11:05

This is what I get.

source_id	type_of_water_source	number_of_people_served
AkKi08081224	shared_tap	3398
AkLu01628224	bio_dirty_well	210
AkRu85234224	tap_in_home_broken	496
HaRu19601224	shared_tap	3322
HaZa21742224	poi_dirty_well	388
SoRu36096224	shared_tap	3786
SoRu37635224	shared_tap	3920
SoRu38776224	shared_tap	3180

11:09

I added a couple of others... Sorry! Well, if you check them you will see which sources have people queueing. The field surveyors also let us know that they measured sources that had queues a few times to see if the queue time changed.

11:11

17

Project instructions format

These **status icons** indicate progress. As we **complete** sections, the status icons turn **green**.

The **current task** we're working on is shown in **blue**, and tasks yet to come are still in red.

The screenshot shows a project interface with a sidebar on the left containing a list of tasks. Each task has a status icon (a circle with a checkmark or a number) and a title. The first task, 'Introduction', has a green checkmark icon. The second task, 'Get to know our data', has a blue circle with the number 1. The third task, 'Dive into sources', has a blue circle with the number 2. The fourth task, 'Unpack the visits', has a blue circle with the number 3. The fifth task, 'Water source quality', has a red circle with the number 4. The sixth task, 'Pollution issues', has a red circle with the number 5. An orange box highlights the first three tasks. The main area on the right shows a chat window with a user named 'Chidi Kunto' who is online. The chat contains two messages. The first message is a dark blue bubble with white text: 'I chose these two: AkRu85234224 HaZa21742224. Ok, so now back to the water_source table. Let's check the records for those source_ids. You can probably remember there is a cool and a "not so cool" way to do it.' The second message is a light blue bubble with black text: 'This is what I get.' followed by a table. The table has three columns: 'source_id', 'type_of_water_source', and 'number_of_people_served'. The table contains six rows of data. The chat window also shows timestamps: 11:05, 11:09, and 11:11.

Introduction
Setting the stage for our data exploration journey.

1 Get to know our data
Exploring the foundational tables and their structure.

2 Dive into sources
Understanding different sources with SELECT.

3 Unpack the visits
Discovering the visit patterns.

4 Water source quality
Understanding water quality.

5 Pollution issues
Correcting pollution data with LIKE and string operations.

Chidi Kunto
Online

I chose these two:
AkRu85234224
HaZa21742224
Ok, so now back to the water_source table. Let's check the records for those source_ids. You can probably remember there is a cool and a "not so cool" way to do it.

11:05

This is what I get.

source_id	type_of_water_source	number_of_people_served
AkKi08081224	shared_tap	3398
AkLu01628224	bio_dirty_well	210
AkRu85234224	tap_in_home_broken	496
HaRu19601224	shared_tap	3322
HaZa21742224	poi_dirty_well	388
SoRu36096224	shared_tap	3786
SoRu37635224	shared_tap	3920
SoRu38776224	shared_tap	3180

11:09

I added a couple of others... Sorry! Well, if you check them you will see which sources have people queueing. The field surveyors also let us know that they measured sources that had queues a few times to see if the queue time changed.

11:11

Project instructions format

The story, tasks, code blocks, and query results are shared by Chidi in this space.

The screenshot shows a project workspace interface. On the left, there is a sidebar with a list of tasks: Introduction, Get to know our data, Dive into sources, Unpack the visits, Water source quality, and Pollution issues. Each task has a number and a status icon. A red box highlights the chat area on the right, which contains a conversation with Chidi Kunto. The chat messages include text, a table of data, and a follow-up message. The table has three columns: source_id, type_of_water_source, and number_of_people_served. The chat area is timestamped 11:05, 11:09, and 11:11.

Chidi Kunto Online

I chose these two:
AkRu85234224
HaZa21742224
Ok, so now back to the water_source table. Let's check the records for those source_ids. You can probably remember there is a cool and a "not so cool" way to do it.

11:05

This is what I get.

source_id	type_of_water_source	number_of_people_served
AkKi08081224	shared_tap	3398
AkLu01628224	bio_dirty_well	210
AkRu85234224	tap_in_home_broken	496
HaRu19601224	shared_tap	3322
HaZa21742224	poi_dirty_well	388
SoRu36096224	shared_tap	3786
SoRu37635224	shared_tap	3920
SoRu38776224	shared_tap	3180

11:09

I added a couple of others... Sorry! Well, if you check them you will see which sources have people queueing. The field surveyors also let us know that they measured sources that had queues a few times to see if the queue time changed.

11:11

Project instructions format

The **timestamps** on these messages are **unique**. We can **reference** these like page numbers when collaborating, or when asking questions.

Chidi Kunto
Online

Introduction
Setting the stage for our data exploration journey.

1 Get to know our data
Exploring the foundational tables and their structure.

2 Dive into sources
Understanding different sources with SELECT.

3 Unpack the visits
Discovering the visit patterns.

4 Water source quality
Understanding water quality.

5 Pollution issues
Correcting pollution data with LIKE and string operations.

I chose these two:
AkRu85234224
HaZa21742224
Ok, so now back to the water_source table. Let's check the records for those source_ids. You can probably remember there is a cool and a "not so cool" way to do it.

11:05

This is what I get.

source_id	type_of_water_source	number_of_people_served
AkKi08081224	shared_tap	3398
AkLu01628224	bio_dirty_well	210
AkRu85234224	tap_in_home_broken	496
HaRu19601224	shared_tap	3322
HaZa21742224	poi_dirty_well	388
SoRu36096224	shared_tap	3786
SoRu37635224	shared_tap	3920
SoRu38776224	shared_tap	3180

11:09

I added a couple of others... Sorry! Well, if you check them you will see which sources have people queueing. The field surveyors also let us know that they measured sources that had queues a few times to see if the queue time changed.

11:11

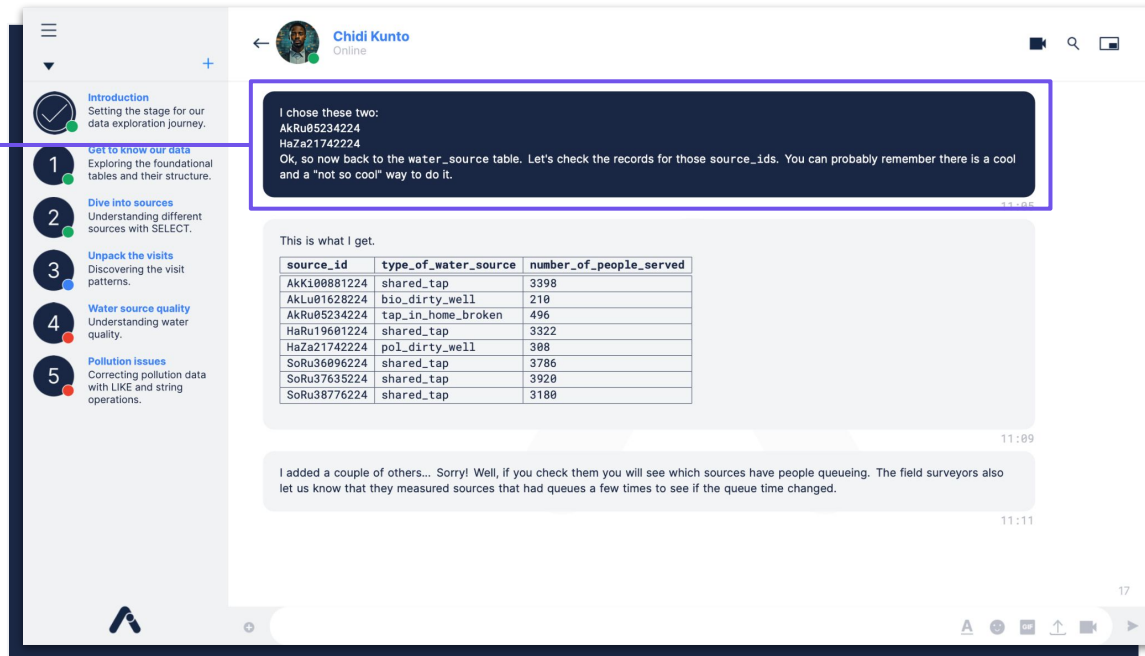
17

Format

The **story** of Maji Ndogo's struggle is told in **grey chat bubbles**.

The **tasks** we should do are **highlighted in dark blue**.

These are the queries we need to create to get the results sets Chidi shares throughout the project.



The screenshot shows a chat application interface. On the left is a sidebar with a list of tasks, each with a numbered icon and a title. The main chat area on the right shows a conversation with a contact named 'Chidi Kunto'. A purple box highlights a specific message in the chat. Below it, a table is displayed, showing data with columns for source_id, type_of_water_source, and number_of_people_served. The chat interface includes a bottom bar with various icons and a page number '17' in the bottom right corner.

Introduction
Setting the stage for our data exploration journey.

1. Get to know our data
Exploring the foundational tables and their structure.

2. Dive into sources
Understanding different sources with SELECT.

3. Unpack the visits
Discovering the visit patterns.

4. Water source quality
Understanding water quality.

5. Pollution issues
Correcting pollution data with LIKE and string operations.

Chidi Kunto
Online

I chose these two:
AkRu85234224
HaZa21742224
Ok, so now back to the water_source table. Let's check the records for those source_ids. You can probably remember there is a cool and a "not so cool" way to do it.

This is what I get.

source_id	type_of_water_source	number_of_people_served
AkKi08081224	shared_tap	3398
AkLu01628224	bio_dirty_well	210
AkRu85234224	tap_in_home_broken	496
HaRu19601224	shared_tap	3322
HaZa21742224	poi_dirty_well	388
SoRu36096224	shared_tap	3786
SoRu37635224	shared_tap	3920
SoRu38776224	shared_tap	3180

I added a couple of others... Sorry! Well, if you check them you will see which sources have people queueing. The field surveyors also let us know that they measured sources that had queues a few times to see if the queue time changed.

Managing SQL queries in MySQL Jupyter notebooks

Jupyter notebooks offer an **interactive environment** that's perfect for **data projects**, especially when working with SQL databases.

We can write SQL queries, execute them, and store the results of those queries – **all in one place**.

We can create notebooks in Jupyter to **organise** our **work**, **summarise** our **findings**, make some **notes**, and **store results** so that we can reference them **later**.

An example notebook is available for Part 1, but we encourage you to **create these on your own** for the rest of the project.

Integrated project notebook

© ExploreAI Academy

This notebook guides us on how to create a notebook for the integrated project.

⚠ This notebook will not run on Google Colab because it cannot connect to a local database. Please make sure that this notebook is running on the same local machine as your MySQL Workbench installation and MySQL `md_water_services` database.

Connecting to our MySQL database

Using our `Access_to_Basic_Services` table in our `united_nations` database we created in MySQL Workbench, we want to answer some questions about our dataset. We can apply the same queries we used in MySQL Workbench in this notebook if we connect to our MySQL server by running the cells below.

```
In [1]: # Load and activate the SQL extension to allow us to execute SQL in a Jupyter notebook.
# If you get an error here, make sure that mysql and pymysql are installed correctly.

%load_ext sql
```

```
In [2]: # Establish a connection to the local database using the '%sql' magic command.
# Replace 'password' with our connection password and 'db_name' with our database name.
# If you get an error here, please make sure the database name or password is correct.

%sql mysql+pymysql://root:password@localhost:3306/md_water_services
```

Managing SQL queries in notebooks

Notebooks are great tools to explore a database. Notebooks contain Python cells:

```
In [ ]: # We run Python commands in here
```

and Markdown cells:

Used to display text with. Running this cell will create formatted text.

We can use Markdown cells to organise our work flow and take notes as we journey through a data project. Below are some useful formatting tools you may need:

Our main goal



To make sure **we're the ones standing out** in an interview, we should be able to **solve any problem** we're given using our SQL skills.



Engaging with this project fully will help you to do that! So forget about the marks, and **build your skills** in SQL.



Several points in this project will be challenging, so we should rely on each other to learn. If you get stuck, **reach out to your teammates** and ask for help.

