# Introduction

## Overview of Produced Water

Produced water, the wastewater generated during the extraction of oil and gas and in other industrial processes, presents a significant environmental challenge. This water typically contains various chemicals and dissolved solids. The oil and gas industry often relies on injection wells to manage this produced water, where it is re-injected underground. While this method is efficient and cost-effective, it carries environmental risks, such as potential groundwater contamination and seismic activity inducement.

In light of these challenges, the Department of Energy (DOE) has funded research to explore treatment options for high-salinity produced water from the oil and gas industry. This research also aims to investigate the feasibility of extracting critical elements, such as lithium, from produced water, thus turning a waste product into a potential resource.

The oil and gas sector faces unique challenges in managing produced water, largely due to its complex chemical composition, high salinity, and the large volumes involved. Various treatment and management techniques are employed, balancing environmental impacts with economic considerations.

## Objectives

This project, aligning with DOE's research interests, is centered around two key objectives:

1. Analyzing the Chemistry of Produced Water: The aim is to understand the composition of produced water and identify elements that lead to scaling in treatment processes, a major impediment in water treatment leading to efficiency losses and increased costs.
2. Evaluating Lithium Availability: We assess the potential for lithium extraction from produced water in the oil and gas industry across the U.S. With the current economic feasibility threshold for lithium extraction set at a concentration of 80 mg/l, this project explores the availability of lithium in produced water from different basins across the US, and anticipates future advancements in extraction technologies that could make it more economically viable.

# Data Acquisition and Initial Processing

## Raw Data Acquisition

The raw dataset for this study was sourced from the United States Geological Survey (USGS) website, available: <https://data.usgs.gov/datacatalog/data/USGS:59d25d63e4b05fe04cc235f9> . This dataset contains extensive information on the chemistry of produced water from various sites. Due to the considerable size of the dataset, it was split into three separate CSV files to facilitate easier handling and uploading.

## Data Loading and Merging

The individual CSV files were loaded into Pandas DataFrames. These separate DataFrames were then concatenated to form a single, comprehensive DataFrame representing the entire dataset. This step was crucial for ensuring a unified analysis across all data points.

# Data Cleaning and Filtering

## Column Removal

The merged DataFrame contained numerous columns, many of which were not pertinent to the specific objectives of this study. To streamline the analysis and focus on relevant data, a list of columns that were not necessary was compiled. These columns included identifiers, location details, physical characteristics, and other parameters that were not directly related to the chemical composition analysis relevant to this study. The unnecessary columns were removed from the DataFrame, resulting in a more focused dataset.

## Data Filtering Based on TDS

The first step in our data cleaning process was filtering out samples with Total Dissolved Solids (TDS) less than 35,000 ppm. This focus on high TDS samples is due to their greater treatment challenges and environmental risks. Produced water from conventional oil and gas sites typically has high salinity, requiring more complex treatment or re-injection into wells. Conversely, low TDS water, often from sources like coalbed methane or geothermal industries, can be treated more easily. Such water can meet standards for reuse within the same industry, other applications, or even be discharged into water bodies, presenting fewer environmental concerns. By doing this filtering, the dataset is narrowed down from 114,944 rows to those that are more relevant for studying high salinity produced water.

## Applying Conditions for Filtering

In the data cleaning and filtering phase, specific criteria were applied to refine the dataset further. These criteria focused on the molar concentration relationships between various elements such as Sodium (Na), Calcium (Ca), Chlorine (Cl), Sulfate (SO4), and Magnesium (Mg). These criteria are based on natural chemical balances typically found in produced water, aiming to identify and exclude samples whose chemical profiles do not align with these natural conditions. Such samples might be unnatural or have problematic analyses and are considered less reliable for the study's purposes.

### *Technical Explanation*

The molar concentration of each element is calculated using its concentration in ppm and its molar mass. The formula used is:

Molarity (M) =

For the analysis, the following molar masses were used:

Sodium (Na): 22.99 g/mol

Calcium (Ca): 40.08 g/mol

Chlorine (Cl): 35.45 g/mol

Sulfate (SO4): 96.06 g/mol

Magnesium (Mg): 24.305 g/mol

Dividing the concentration of each element by its respective molar mass provides the molarity, which is a basis for comparison across different water samples.

### *Natural Conditions for Produced Water*

A sample is considered representative of natural produced water if it typically meets the following conditions:

Molar Sodium (Na) greater than Molar Calcium (Ca): Reflecting the ion balance in natural water bodies where sodium is often more concentrated than calcium.

Molar Chlorine (Cl) greater than Molar Sulfate (SO4): Based on the common occurrence of chloride as a dominant anion in many natural waters, particularly those associated with oil and gas production.

Molar Calcium (Ca) greater than half of Molar Magnesium (Mg): Indicative of the natural geochemical processes affecting water composition.

Applying these conditions helps filter the dataset to focus on samples that reflect the expected natural chemical composition of produced water. This step is crucial for ensuring the accuracy and reliability of the subsequent analyses.

The dataset, post-application of these conditions, offers a more accurate representation of natural produced water chemistry, essential for the objectives of the study.

### *Charge Balance Criteria*

For the dataset to be scientifically valid and reflective of naturally occurring water chemistry, it was decided that only those samples with a USGS charge balance within the range of -10 to +10 would be retained. This range was chosen as it is generally accepted in geochemical studies as indicative of a balanced and realistic chemical composition. Samples outside this range could suggest potential inaccuracies in the data or unusual water chemistry, which might not align with the study's objectives. To apply this criterion, the dataset was filtered to exclude any rows where the 'chargebalance' value did not fall within the specified range of -10 to +10.

## Handling Missing Data

Different strategies were employed to address missing values in various chemical constituents of the produced water.

### *Filling NaN Values for Basic Elements*

For key elements such as Potassium and Sodium (KNa), Potassium (K), Sodium (Na), Calcium (Ca), Chlorine (Cl), Sulfate (SO4), and Magnesium (Mg), missing values (NaN) were filled with zeros. This approach was necessary for calculation purposes, allowing for further processing and analysis without data loss.

### *Calculating Sodium (Na) Values Where Missing*

A more nuanced approach was taken for Sodium (Na) due to its importance and relationship with other elements. The calculation for Na was based on the presence or absence of KNa and K:

* If Na was missing but both KNa and K were present, Na was populated with the difference between KNa and K.
* If Na was missing, KNa was present, but K was not, Na was set equal to KNa.

This conditional approach ensured a more accurate representation of the Na values in the dataset. Rows where Na or Cl were still missing after these calculations were removed from the dataset, as these elements are crucial for the study.

### *Handling Carbonates*

For Carbonate (CO3) and Bicarbonate (HCO3), missing values were also treated with specific strategies:

* Missing CO3 values were replaced with zeros.
* Missing HCO3 values were replaced with available Alkalinity as HCO3 (ALKHCO3) values where possible.
* In cases where both HCO3 and ALKHCO3 were missing, HCO3 was calculated based on the difference between cations and anions, divided by the molar mass of HCO3.

These steps were essential to maintain the chemical balance in the dataset and to ensure that carbonate chemistry was accurately represented.

Each of these strategies for handling missing data was chosen to respect the chemical properties and interactions of the elements involved, ensuring that the dataset remained scientifically valid and robust for subsequent analyses. The updated datasets, post these cleaning steps, were saved as df\_filtered\_Na\_Cl.csv and df\_filtered\_estimated\_HCO3.csv for further use in the study.

# Visualization

## Methodology for Generating Box and Violin Plots

The geological basins under study are presented to the user. Users are prompted to select the basins of interest by entering the corresponding numbers. The input is processed to create a list of selected basins. The main DataFrame is filtered to include only the data from the basins selected by the user.

The elements chosen for concentration analysis are Calcium (Ca), Magnesium (Mg), Bicarbonate (HCO3), Silicon (Si), Total Iron (FeTot), Barium (Ba), Strontium (Sr), and Lithium (Li). For each selected element, a box plot is generated. Two functions are defined for generating the box plots and violin plots. The plots display the distribution of the element's concentration across the selected basins. The figure size is dynamically determined based on the number of basins selected to ensure clear visualization.

Both box and violin plots are titled with the name of the element being analyzed. The X-axis represents the basin categories, and the Y-axis shows the concentration of the element in mg/L. The plots are saved to a designated directory for further use and reference. The code includes checks to ensure that the selected element exists in the DataFrame. If an element is not found, a message is displayed, and the plot for that element is not generated.

### Tools and Libraries Used

* Pandas for data manipulation.
* Matplotlib and Seaborn for data visualization.

The generated plots provide a visual representation of the variability and distribution of elemental concentrations across different geological basins, aiding in comparative analysis and interpretation of geochemical data.

## Methodology for Mapping Lithium Concentration Clusters

The objective of this visualization is to spatially visualize the distribution of lithium (Li) concentrations in various geological basins using a cluster mapping approach with a focus on both overall concentration clusters and high concentration areas using interactive maps.

The DataFrame is cleaned to remove NaN values specifically from the 'LATITUDE', 'LONGITUDE', and 'Li' columns. This ensures that only complete records are used for the analysis. A function, assign\_cluster, is created to categorize lithium concentrations into distinct clusters based on predefined concentration ranges (<= 20 ppm, 21-40 ppm, 41-60 ppm, 61-80 ppm, > 80 ppm). This categorization aids in differentiating areas based on lithium concentration levels.

The assign\_cluster function is applied to the 'Li' column of the DataFrame. This process assigns each record to a lithium concentration cluster, creating a new column 'Li\_Cluster'. A GeoDataFrame is created from the cleaned DataFrame. The 'LATITUDE' and 'LONGITUDE' columns are used to create point geometries, facilitating spatial analysis and visualization. A Folium map object, m, is initialized, centered around the mean latitude and longitude of the data points to provide a comprehensive view of the study area.

A color scheme is defined for the different lithium concentration clusters, aiding in visual differentiation of the clusters on the map. Each data point, representing a geographical location with a specific lithium concentration, is plotted on the map as a circle marker. The color of the marker corresponds to its assigned lithium concentration cluster. An HTML-based legend is created to provide context for the color scheme used in the map. This legend is integrated into the Folium map, enhancing the interpretability of the visualization. The final map, complete with data points and the legend, is saved as an HTML file. This file can be viewed in a web browser, allowing for interactive exploration of the data.

For mapping high lithium concentrations with gradient color scale, the DataFrame is filtered to include only records where lithium concentrations are equal to or greater than 80 mg/L. This subset focuses the analysis on areas with notably high lithium levels. A GeoDataFrame for the high concentration data is created. Then a linear color scale ranging from medium red to dark red is developed to represent varying levels of high lithium concentrations. The high concentration data points on a separate Folium map with colors corresponding to their specific lithium levels are plotted.

Both maps are designed to be interactive, allowing for detailed exploration of lithium concentrations.

### Tools and Libraries Used

* Pandas for DataFrame operations.
* Geopandas for handling geospatial data.
* Folium for creating interactive maps.

This methodology enables a clear and interactive visualization of lithium concentration distributions across geological basins. It provides valuable insights into spatial patterns and concentration clusters, which are essential for geochemical analysis and decision-making in resource exploration and environmental studies.

# Conclusion

This project sheds light on the chemistry of produced water in the oil and gas industry, offering practical solutions for environmental and resource management. The key achievements include:

* *Detailed Analysis of Produced Water:* Understanding the complex makeup of produced water is crucial for better treatment strategies. This project provides clear insights into what's in this water, which is vital for companies looking to manage it more effectively.
* *Lithium Extraction Potential:* Investigating how to extract lithium from produced water could turn an environmental problem into a valuable resource. With lithium's growing demand, especially in renewable energy, this could be a game-changer.
* *Alignment with DOE Goals:* This study aligns with the Department of Energy’s objectives, contributing to research that can lead to more sustainable industry practices.
* *Practical Methodologies:* The approach taken in this study – from gathering and processing data to analyzing chemical content – is straightforward and can be replicated in similar studies or applied in real-world scenarios.
* *Support for Better Water Management:* The findings support the development of more sustainable water management practices, helping companies reduce environmental impacts while finding value in what was once waste.
* *Opens Doors for Resource Recovery:* The study encourages further exploration into recovering resources from produced water, promoting innovative approaches in the industry.

In summary, this project not only offers a deeper understanding of produced water's composition but also paves the way for turning environmental challenges into economic opportunities. The insights gained could be significant for stakeholders looking to enhance environmental responsibility while exploring new avenues for resource utilization.