BIG DATA ANALYSIS USING IBM CLOUD

PHASE 5: DOCUMENTATION AND PRESENTING

TOPIC: SOCIAL MEDIA ANALYSIS

Design thinking and development phases for big data analysis using cloud computing involve a structured approach to identifying, designing, and implementing data analysis solutions that leverage the scalability, flexibility, and cost-effectiveness of cloud services. Below are the key phases and steps for such a project:

1. Define the Problem:

- Identify the specific problem or business goal you want to address with big data analysis.

- Understand the data sources, volume, and variety involved.

- Define success criteria and key performance indicators (KPIs)

2. Empathize and Research:

- Conduct interviews and surveys with stakeholders to understand their needs and pain points.

- Research industry trends, best practices, and competitors' approaches.

- Identify potential data sources and cloud platforms that can support your objectives.

3. Ideation:

- Brainstorm and generate ideas for data analysis solutions.

- Prioritize and select the most promising concepts that align with the project's objectives.

- Consider data storage and processing technologies that can handle big data efficiently.

4. Prototyping:

- Create prototypes or proof-of-concept (PoC) solutions using cloud-based tools and services.

- Develop small-scale models to test your ideas and evaluate their feasibility.

- Validate the technical aspects of your design and its potential impact on the problem.

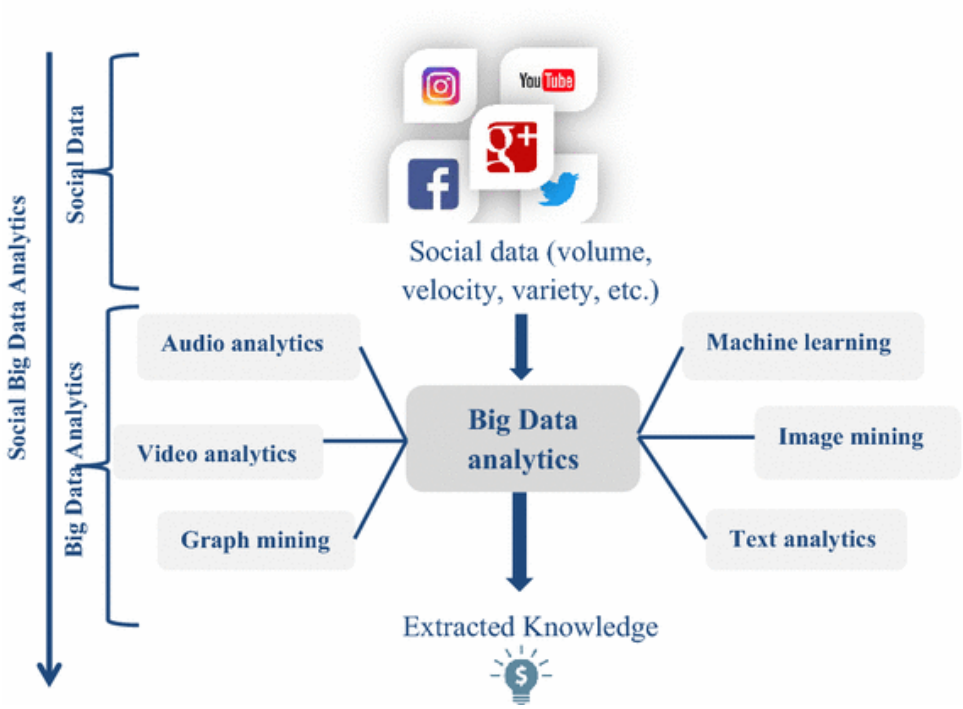
5. Define Data Requirements:

- Identify the data sources required to address the problem.

- Determine data volume, velocity, variety, and veracity.

- Define data integration and data transformation needs.

Social media analytics using big data analysis data architecture:



Understanding Social Media Big Data:

The term "big data" in social media refers to the enormous amount of structured and unstructured data generated by users. This data encompasses text, images, videos, comments, likes, shares, and much more. The scale and complexity of this data require advanced analytics and processing methods, making it a prime candidate for big data analysis.

Key Applications:

1. Sentiment Analysis:

- One of the primary applications of big data analysis in social media is sentiment analysis. It involves evaluating public sentiment toward brands, products, or topics. By analyzing social media posts and comments, businesses can gauge customer satisfaction and respond to issues in real time.

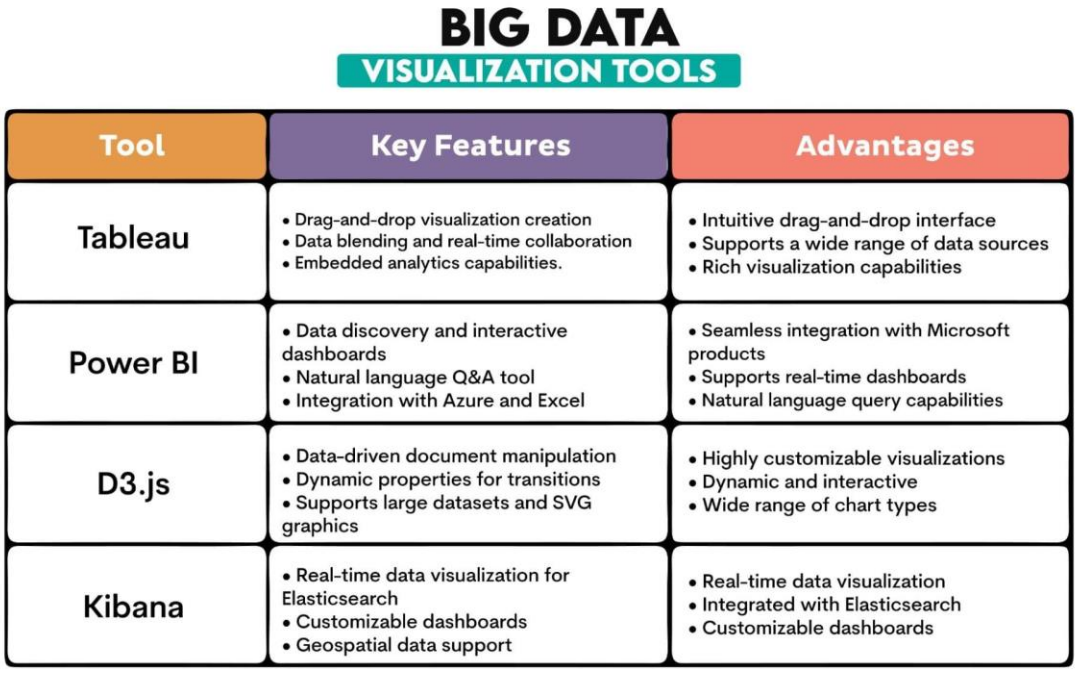
2. Trend Identification:

- Social media platforms are a hub for emerging trends. Big data analysis can help identify these trends, enabling businesses to adapt their strategies, launch relevant products, or engage in timely marketing campaigns.

3. Customer Insights:

- Big data analysis helps businesses understand their customers on a deeper level. By analyzing user behavior, preferences, and interactions, companies can tailor their offerings and marketing efforts to a specific audience.

BIG DATA VISUALIZATION TOOLS :



Tool	Key Features	Advantages
Tableau	• Drag-and-drop visualization creation • Data blending and real-time collaboration • Embedded analytics capabilities.	• Intuitive drag-and-drop interface • Supports a wide range of data sources • Rich visualization capabilities
Power BI	• Data discovery and interactive dashboards • Natural language Q&A tool • Integration with Azure and Excel	• Seamless integration with Microsoft products • Supports real-time dashboards • Natural language query capabilities
D3.js	• Data-driven document manipulation • Dynamic properties for transitions • Supports large datasets and SVG graphics	• Highly customizable visualizations • Dynamic and interactive • Wide range of chart types
Kibana	• Real-time data visualization for Elasticsearch • Customizable dashboards • Geospatial data support	• Real-time data visualization • Integrated with Elasticsearch • Customizable dashboards

Visualization plays a critical role in big data analysis as it helps make complex datasets more understandable, revealing insights and patterns that might be hidden in the raw data. Here are some common visualization techniques used in big data analysis:

1. Bar Charts and Histograms

- Bar charts are useful for showing the distribution of categorical data, such as the frequency of different categories in a dataset.

- Histograms are a type of bar chart used to display the distribution of continuous numerical data.

2. Scatter Plots:

- Scatter plots are used to display the relationship between two continuous variables. Each point represents an observation, and the position of the point on the chart indicates the values of the two variables.

3. Line Charts:

- Line charts are effective for showing trends and changes over time. They are commonly used in time-series analysis to visualize data over sequential time intervals.

4. Heatmaps:

- Heatmaps are used to represent data in a two-dimensional matrix, where colors represent the values of individual cells. They are valuable for visualizing correlations, clustering, and patterns in large datasets.

5. Box Plots:

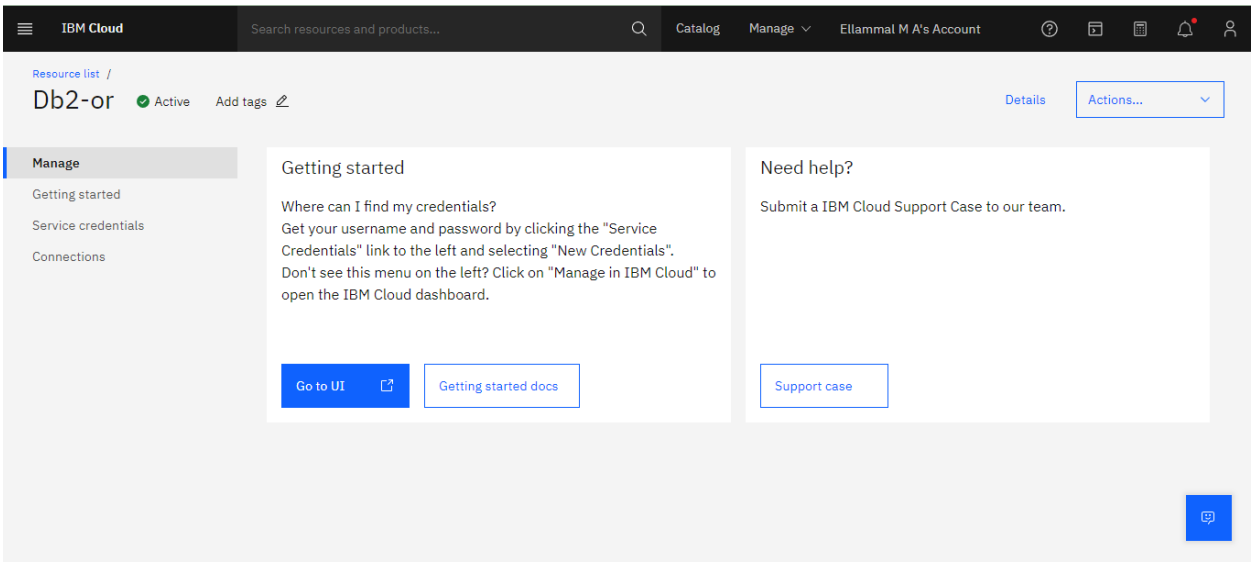
- Box plots (box-and-whisker plots) are used to display the distribution of numerical data, providing information about the median, quartiles, and potential outliers.

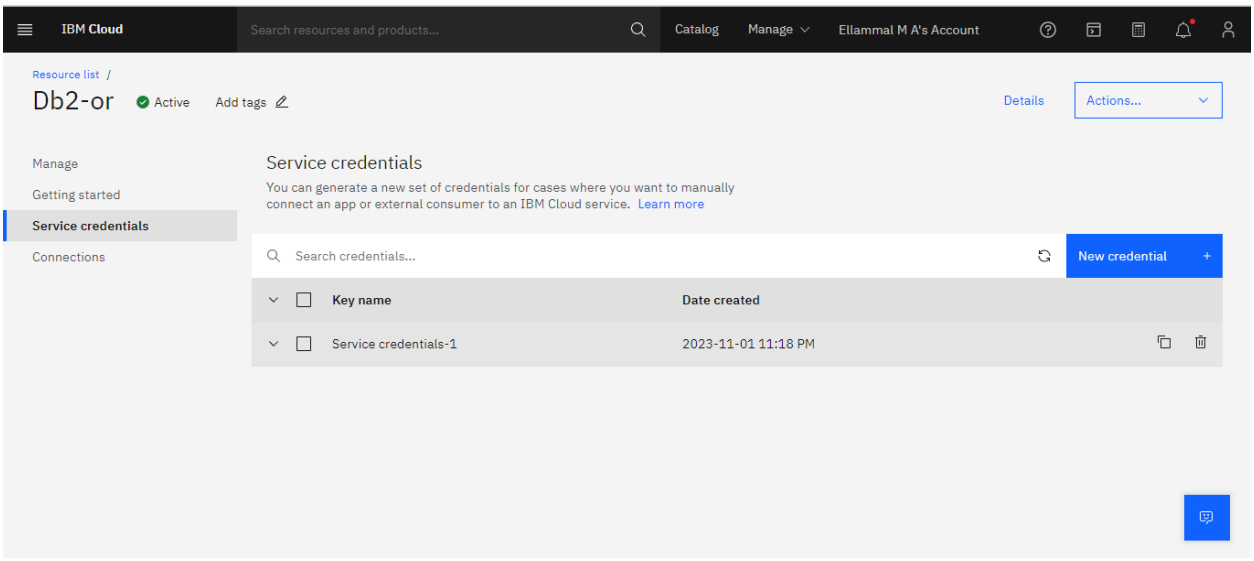
PYTHON CODE :

```
import tweepy
import pandas as pd
import matplotlib.pyplot as plt

# Set up Twitter API credentials
consumer_key = 'your_consumer_key'
consumer_secret = 'your_consumer_secret'
access_token = 'your_access_token'
access_token_secret = 'your_access_token_secret'

# Authenticate with Twitter API
auth = tweepy.OAuthHandler(consumer_key, consumer_secret)
auth.set_access_token(access_token, access_token_secret)

api = tweepy.API(auth)

# Define the search query and number of tweets to retrieve
search_query = 'data science'
num_tweets = 100

# Collect tweets
tweets = tweepy.Cursor(api.search, q=search_query,
lang='en').items(num_tweets)
```

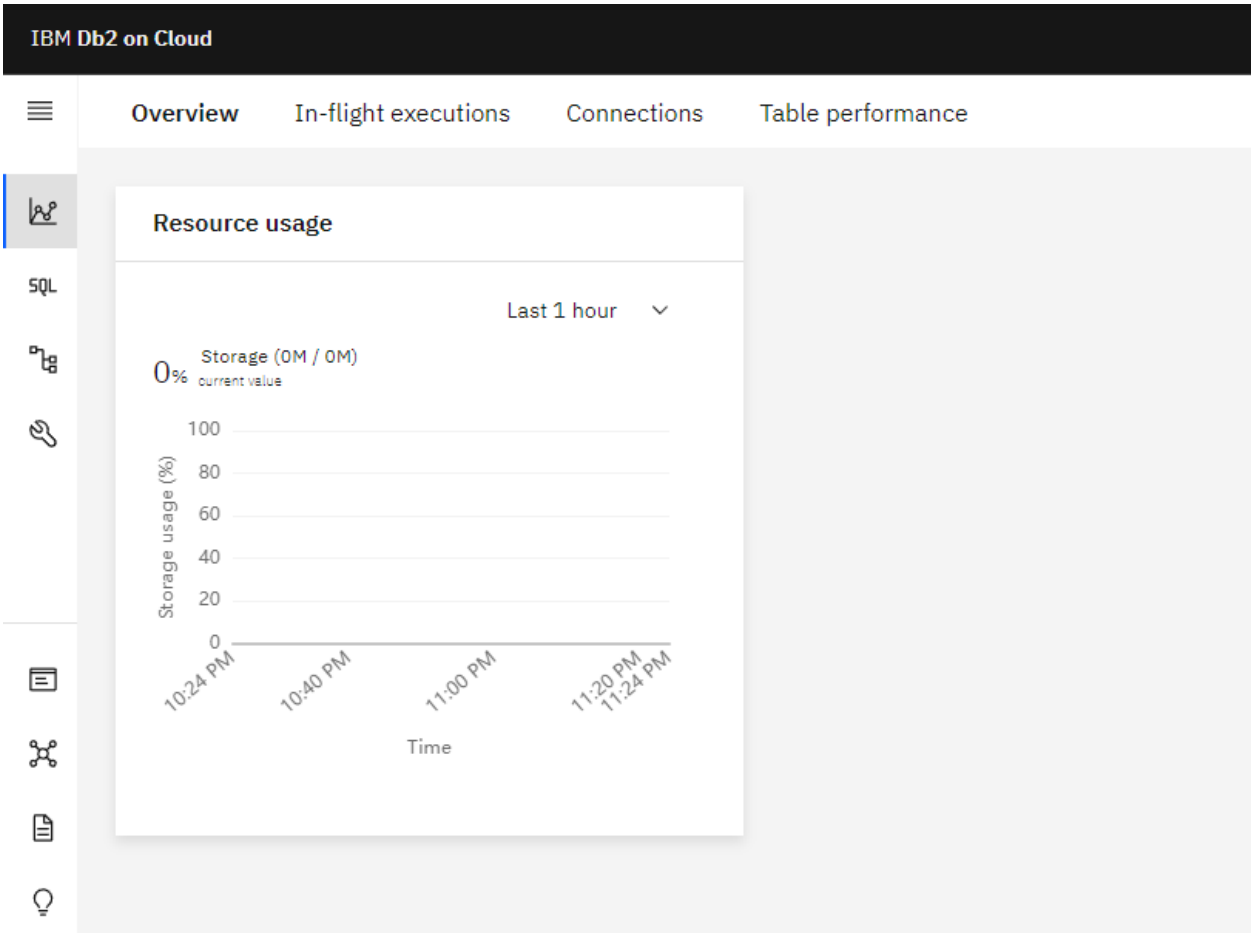
```
# Create a list to store tweet data
tweet_data = []

for tweet in tweets:
    tweet_data.append({
        'Text': tweet.text,
        'User': tweet.user.screen_name,
        'Retweets': tweet.retweet_count,
        'Favorites': tweet.favorite_count,
        'Date': tweet.created_at
    })

# Create a DataFrame from the tweet data
df = pd.DataFrame(tweet_data)

# Analyze the data (e.g., plot a histogram of retweets)
plt.hist(df['Retweets'], bins=10)
plt.xlabel('Number of Retweets')
plt.ylabel('Number of Tweets')
plt.title('Retweet Distribution')
plt.show()

# Analyze the data (e.g., display the top 10 tweets by favorites)
top_tweets = df.nlargest(10, 'Favorites')
print('Top 10 Tweets by Favorites:')
print(top_tweets[['User', 'Favorites', 'Text']])
```

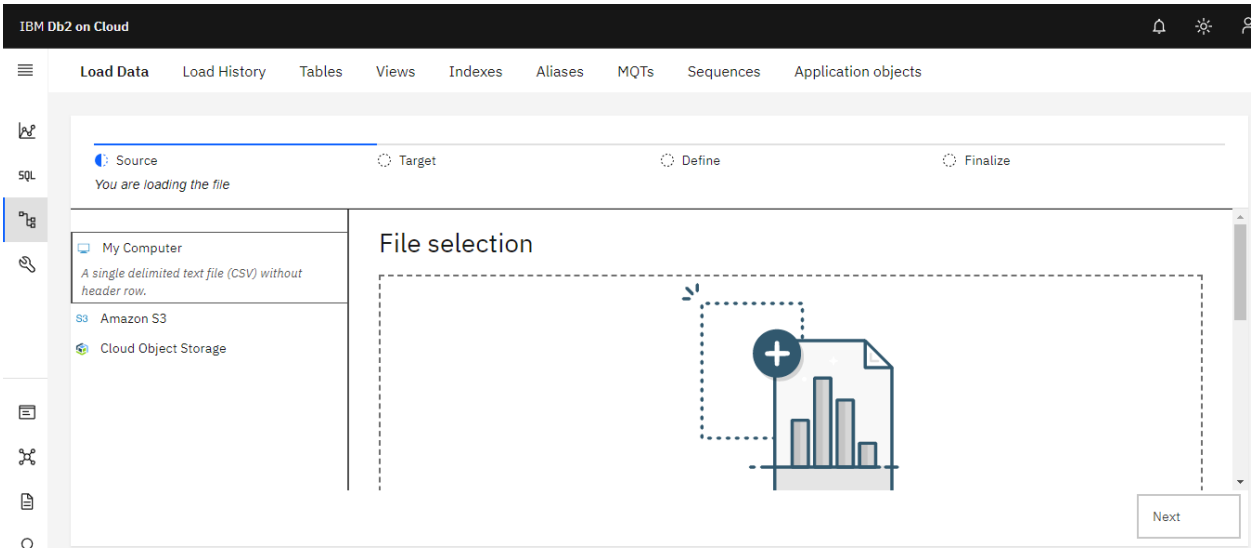
PHASE 3

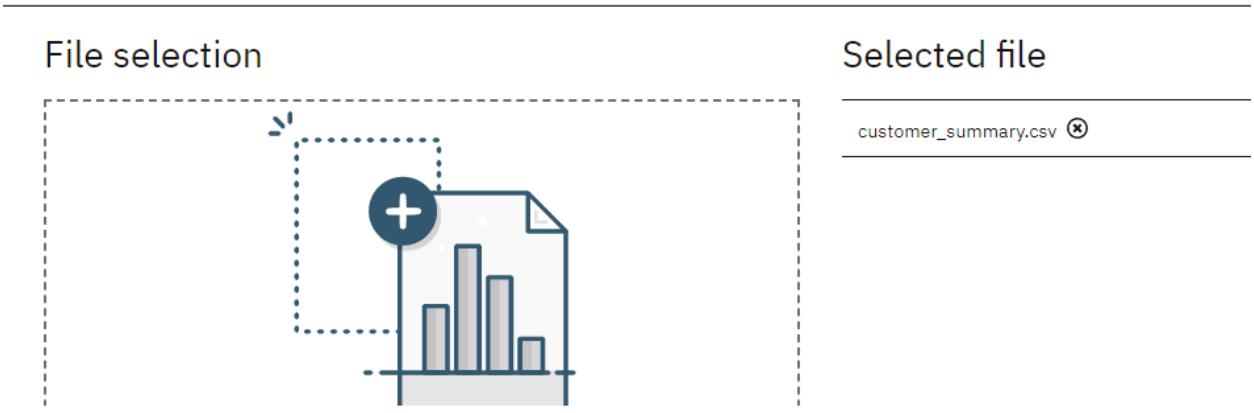


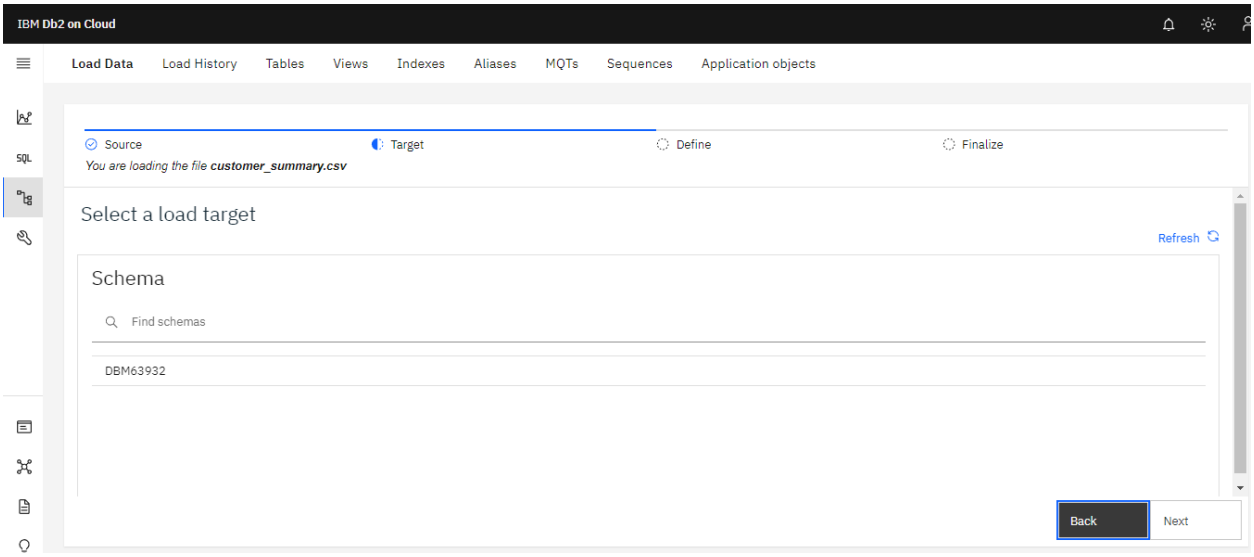


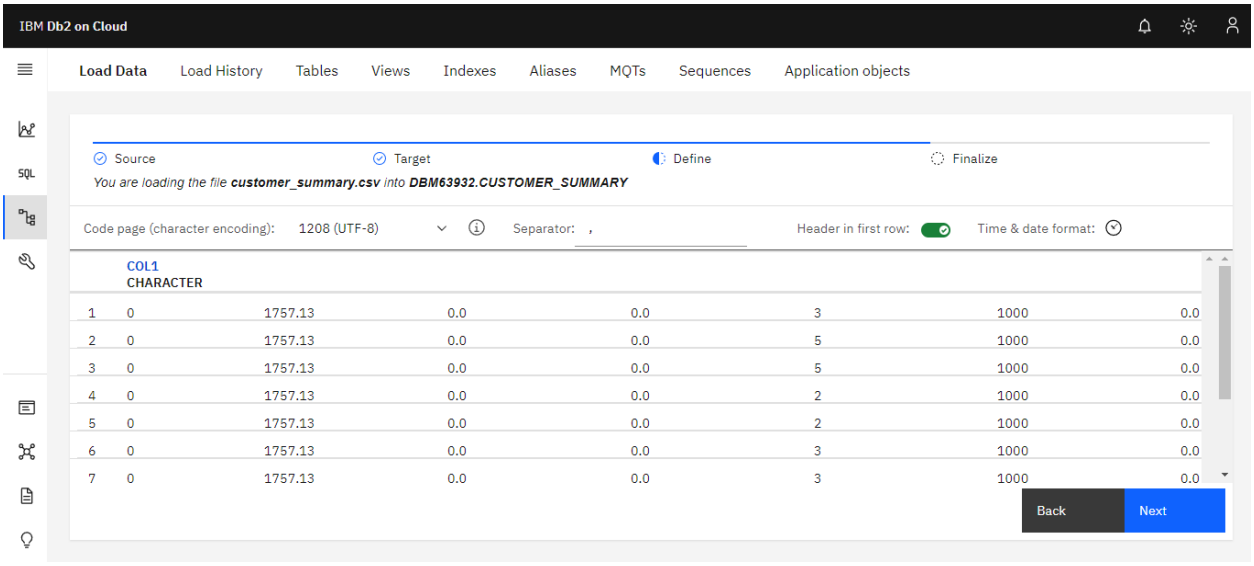

```
"-u",
        "dbm63932",
        "-p",
        "W3KPfJv9swuf699o",
        "--ssl",
        "--sslCAFile",
        "1dd14d0c-1b52-4f63-a606-53ecba28771d",
        "--authenticationDatabase",
        "admin",
        "--host",
```

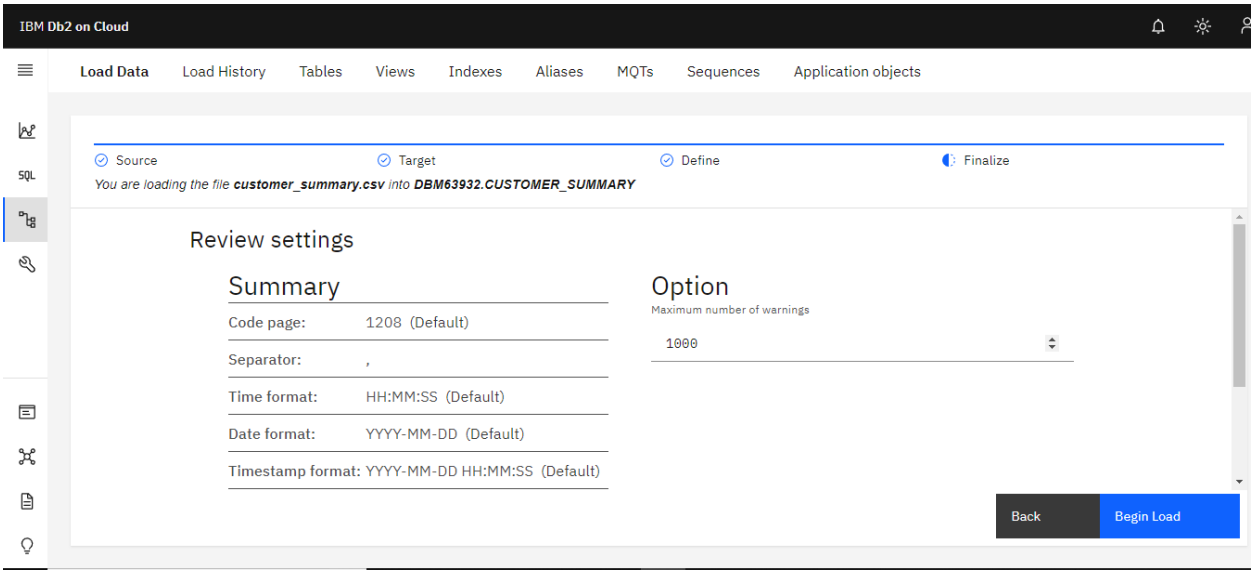

IBM Db2 on Cloud
Overview
In-flight executions
Connections
Table performance
SQL
Resource usage
Last 1 hour
0%
Storage (0M / 0M)
current value
100
80
60
40
20
0
Storage usage (%)
10:24 PM
10:40 PM
11:00 PM
11:20 PM
11:24 PM
Time

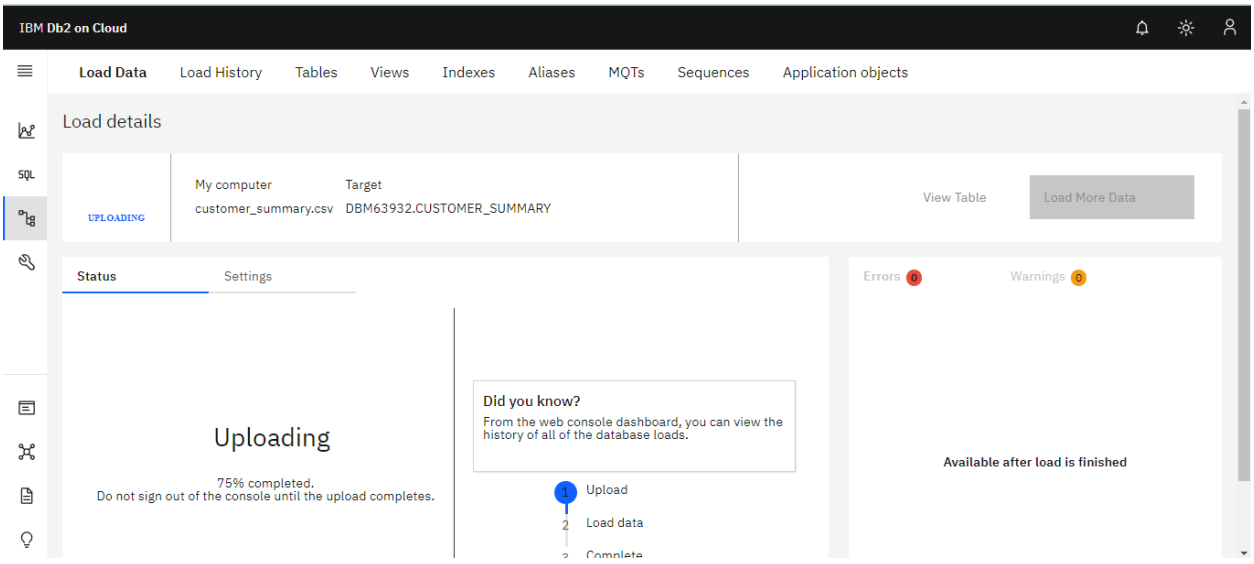

IBM Db2 on Cloud
Load Data
Load History
Tables
Views
Indexes
Aliases
MQTs
Sequences
Application objects
SQL
Source
Target
Define
Finalize
You are loading the file
My Computer
A single delimited text file (CSV) without header row.
Amazon S3
Cloud Object Storage
File selection
Next

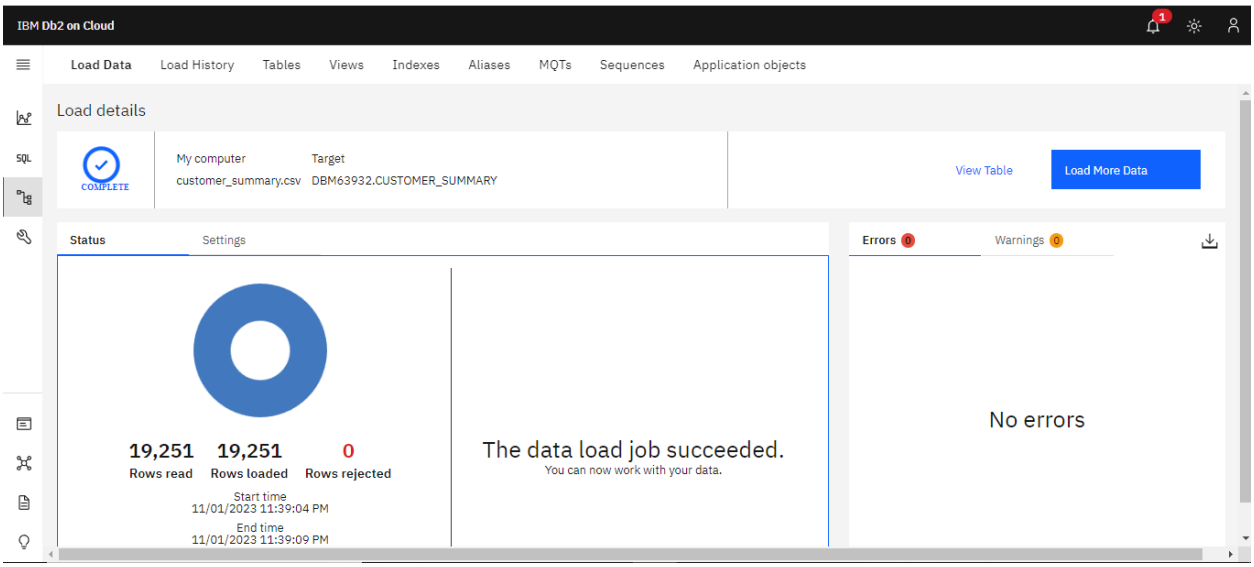

File selection
Selected file
customer_summary.csv

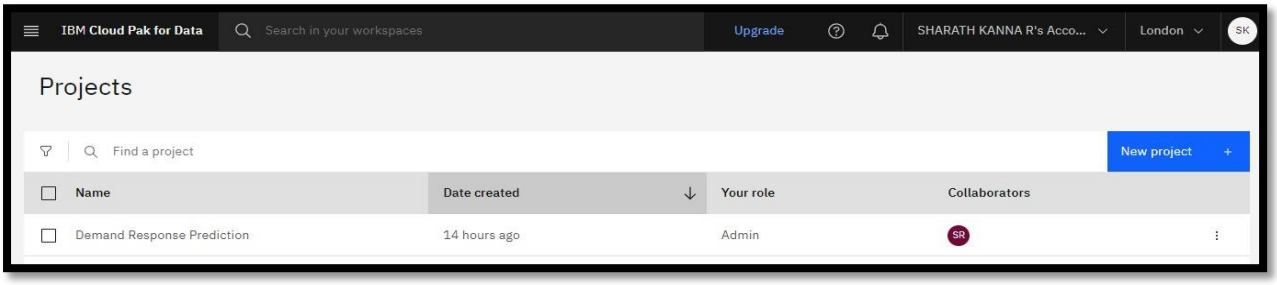

IBM Db2 on Cloud
Load Data
Load History
Tables
Views
Indexes
Aliases
MQTs
Sequences
Application objects
Source
Target
Define
Finalize
You are loading the file customer_summary.csv
Select a load target
Refresh
Schema
Find schemas
DBM63932
Back
Next

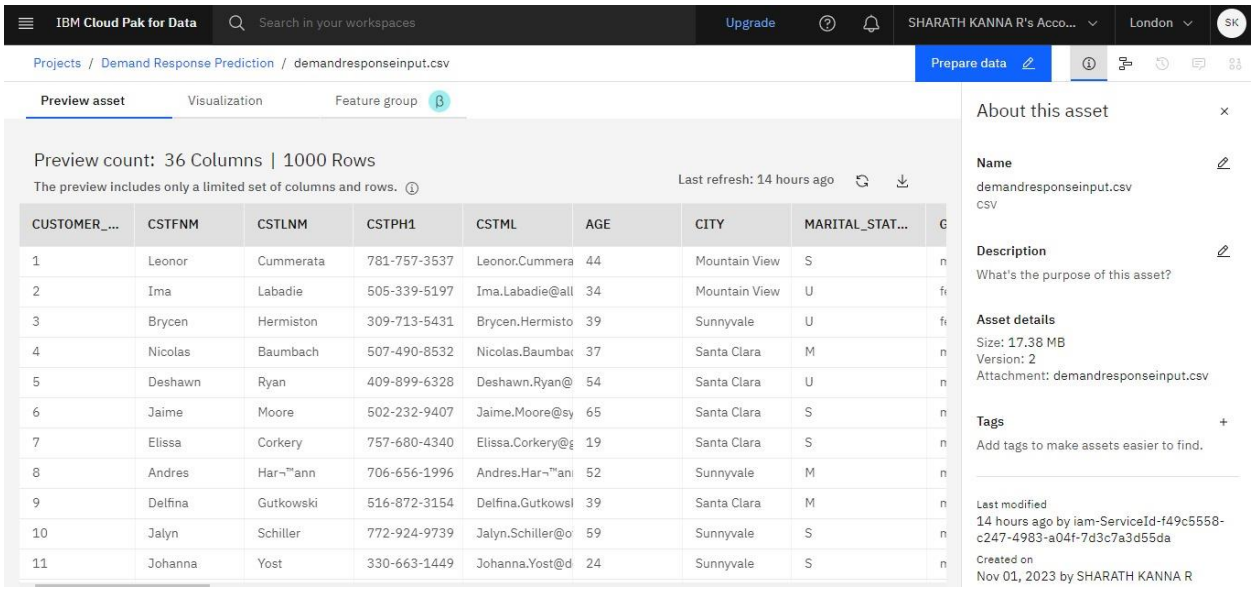

IBM Db2 on Cloud
Load Data
Load History
Tables
Views
Indexes
Aliases
MQTs
Sequences
Application objects
Source
Target
Define
Finalize
You are loading the file customer_summary.csv into DBM63932.CUSTOMER_SUMMARY
Code page (character encoding):
1208 (UTF-8)
Separator:
,
Header in first row:
Time & date format:
COL1
CHARACTER
1 0 1757.13 0.0 0.0 3 1000 0.0
2 0 1757.13 0.0 0.0 5 1000 0.0
3 0 1757.13 0.0 0.0 5 1000 0.0
4 0 1757.13 0.0 0.0 2 1000 0.0
5 0 1757.13 0.0 0.0 2 1000 0.0
6 0 1757.13 0.0 0.0 3 1000 0.0
7 0 1757.13 0.0 0.0 3 1000 0.0
Back
Next


IBM Db2 on Cloud
Load Data
Load History
Tables
Views
Indexes
Aliases
MQTs
Sequences
Application objects
Source
Target
Define
Finalize
You are loading the file customer_summary.csv into DBM63932.CUSTOMER_SUMMARY
Review settings
Summary
Code page: 1208 (Default)
Separator: ,
Time format: HH:MM:SS (Default)
Date format: YYYY-MM-DD (Default)
Timestamp format: YYYY-MM-DD HH:MM:SS (Default)
Option
Maximum number of warnings
1000
Back
Begin Load


IBM Db2 on Cloud
Load Data
Load History
Tables
Views
Indexes
Aliases
MQTs
Sequences
Application objects
Load details
UPLOADING
My computer
customer_summary.csv
Target
DBM63932.CUSTOMER_SUMMARY
View Table
Load More Data
Status
Settings
Errors 0
Warnings 0
Uploading
75% completed.
Do not sign out of the console until the upload completes.
Did you know?
From the web console dashboard, you can view the history of all of the database loads.
1 Upload
2 Load data
Available after load is finished


IBM Db2 on Cloud
Load Data
Load History
Tables
Views
Indexes
Aliases
MQTs
Sequences
Application objects
Load details
COMPLETE
My computer
customer_summary.csv
Target
DBM63932.CUSTOMER_SUMMARY
View Table
Load More Data
Status
Settings
Errors 0
Warnings 0
19,251
Rows read
19,251
Rows loaded
0
Rows rejected
Start time
11/01/2023 11:39:04 PM
End time
11/01/2023 11:39:09 PM
The data load job succeeded.
You can now work with your data.
No errors


IBM Cloud Pak for Data
Search in your workspaces
Upgrade
SHARATH KANNA R's Acco...
London
SK
Projects
Find a project
New project
Name
Date created
Your role
Collaborators
Demand Response Prediction
14 hours ago
Admin
SR


IBM Cloud Pak for Data
Search in your workspaces
Upgrade
SHARATH KANNA R's Acco...
London
SK
Projects / Demand Response Prediction / demandresponseinput.csv
Prepare data
Preview asset
Visualization
Feature group
Preview count: 36 Columns | 1000 Rows
The preview includes only a limited set of columns and rows.
Last refresh: 14 hours ago
CUSTOMER_...
CSTFNM
CSTLNM
CSTPH1
CSTML
AGE
CITY
MARITAL_STAT...
1 Leonor Cummerata 781-757-3537 Leonor.Cummera 44 Mountain View S
2 Ima Labadie 505-339-5197 Ima.Labadie@all 34 Mountain View U
3 Brycen Hermiston 309-713-5431 Brycen.Hermisto 39 Sunnyvale U
4 Nicolas Baumbach 507-490-8532 Nicolas.Baumbac 37 Santa Clara M
5 Deshawn Ryan 409-899-6328 Deshawn.Ryan@ 54 Santa Clara U
6 Jaime Moore 502-232-9407 Jaime.Moore@sy 65 Santa Clara S
7 Elissa Corkery 757-680-4340 Elissa.Corkery@g 19 Santa Clara S
8 Andres Har¬™ann 706-656-1996 Andres.Har¬™an 52 Sunnyvale M
9 Delfina Gutkowski 516-872-3154 Delfina.Gutkowsl 39 Santa Clara M
10 Jalyn Schiller 772-924-9739 Jalyn.Schiller@o 59 Sunnyvale S
11 Johanna Yost 330-663-1449 Johanna.Yost@d 24 Sunnyvale S
About this asset
Name
demandresponseinput.csv
CSV
Description
What's the purpose of this asset?
Asset details
Size: 17.38 MB
Version: 2
Attachment: demandresponseinput.csv
Tags
Add tags to make assets easier to find.
Last modified
14 hours ago by iam-ServiceId-f49c5558-c247-4983-a04f-7d3c7a3d55da
Created on
Nov 01, 2023 by SHARATH KANNA R

IBM Cloud Pak for Data | Search in your workspaces | Upgrade | SHA

Projects / Demand Response Prediction / demandresponseinput.csv

Preview asset | Visualization | Feature group β

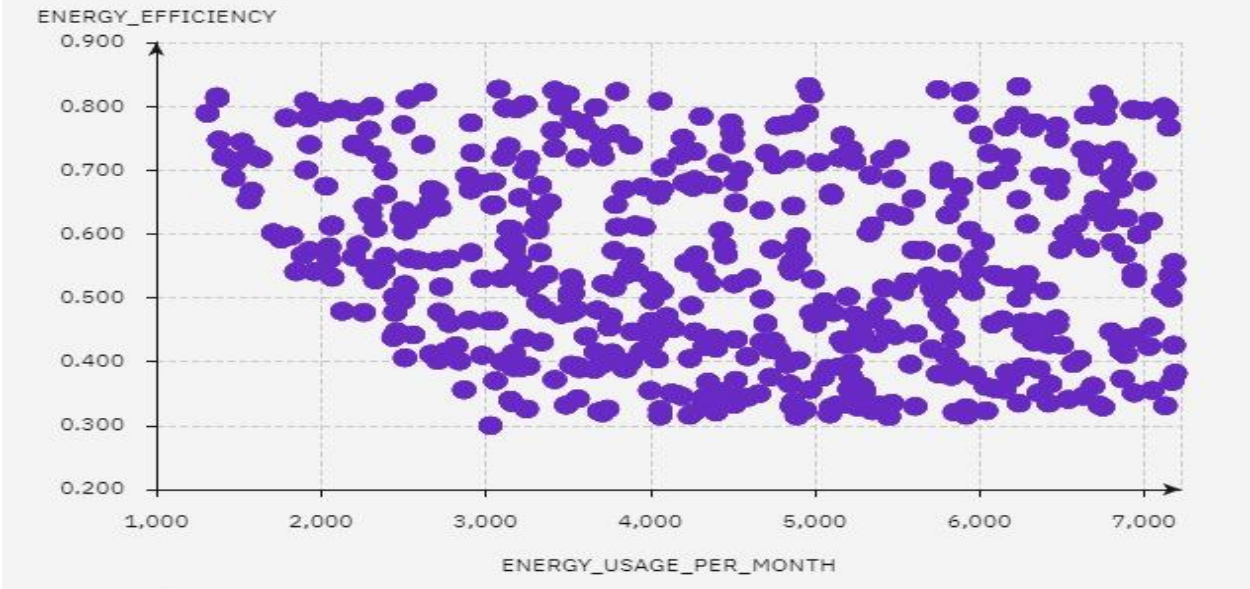
Preview count: 36 Columns | 1000 Rows

The preview includes only a limited set of columns and rows.

Last refresh: 14 hours ago

CUSTOMER_...	CSTFNM	CSTLNM	CSTPH1	CSTML	AGE	CITY	MARITAL_STAT...
1	Leonor	Cummerata	781-757-3537	Leonor.Cummera	44	Mountain View	S
2	Ima	Labadie	505-339-5197	Ima.Labadie@all	34	Mountain View	U
3	Brycen	Hermiston	309-713-5431	Brycen.Hermisto	39	Sunnyvale	U
4	Nicolas	Baumbach	507-490-8532	Nicolas.Baumbac	37	Santa Clara	M
5	Deshawn	Ryan	409-899-6328	Deshawn.Ryan@	54	Santa Clara	U
6	Jaime	Moore	502-232-9407	Jaime.Moore@sy	65	Santa Clara	S
7	Elissa	Corkery	757-680-4340	Elissa.Corkery@g	19	Santa Clara	S
8	Andres	Har¬™ann	706-656-1996	Andres.Har¬™an	52	Sunnyvale	M
9	Delfina	Gutkowski	516-872-3154	Delfina.Gutkowsl	39	Santa Clara	M
10	Jalyn	Schiller	772-924-9739	Jalyn.Schiller@o	59	Sunnyvale	S
11	Johanna	Yost	330-663-1449	Johanna.Yost@d	24	Sunnyvale	S

CHART TYPE | Scatter plot | Line | Multi-series | ACTIONS



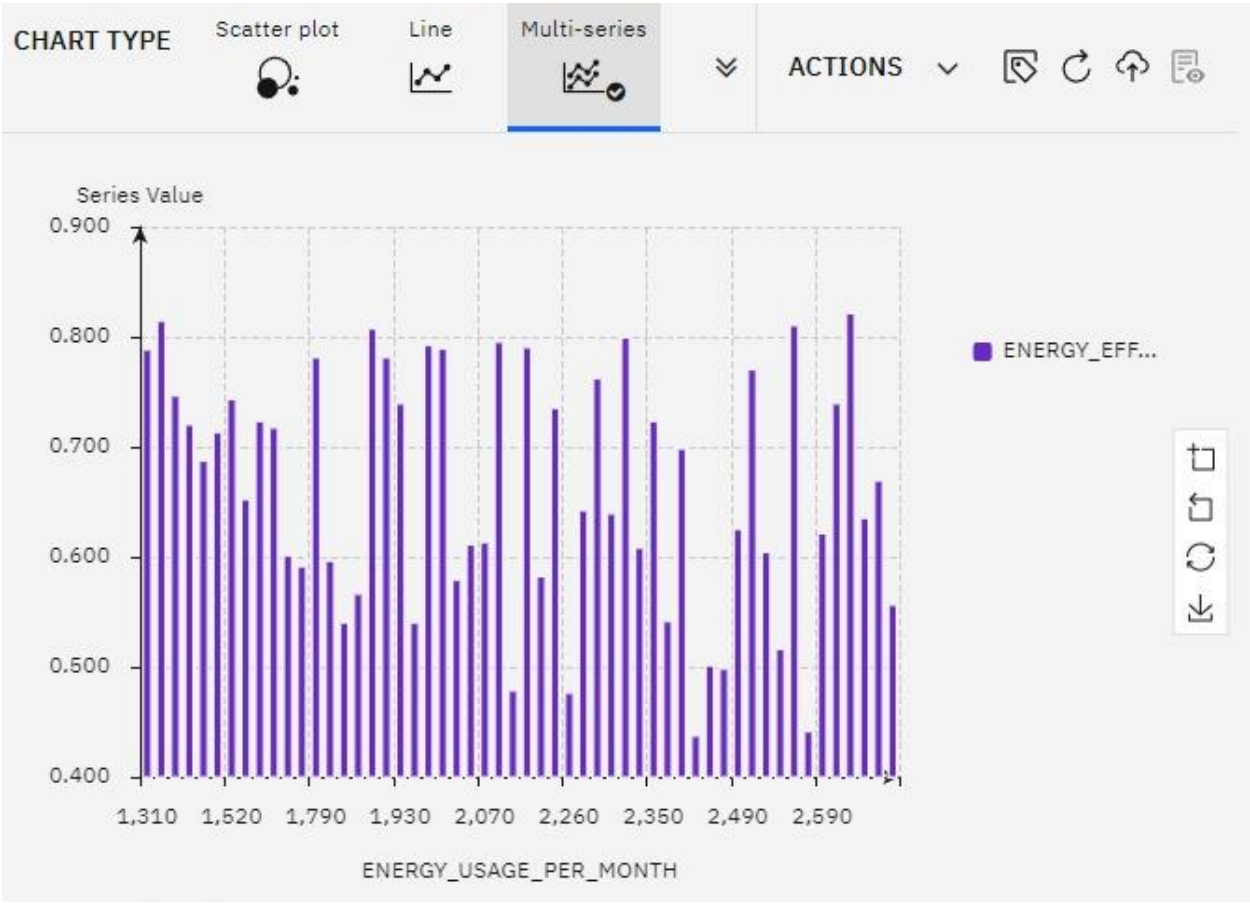


CHART TYPE
Scatter plot
Line
Multi-series
ACTIONS
Series Value
0.900
0.800
0.700
0.600
0.500
0.400
1,310
1,520
1,790
1,930
2,070
2,260
2,350
2,490
2,590
ENERGY_USAGE_PER_MONTH
ENERGY_EFF...