Data Science Infrastructure (DSI) Project

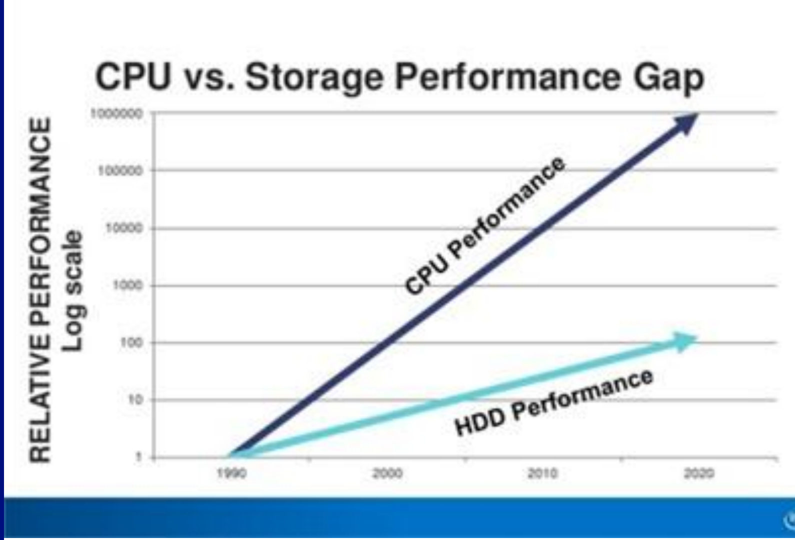
Giovanni Michel

08/04/2022

Mentors Jesus Pulido
Terry Turton

Motivation for DSI

1. Inspired by the time to gather and run simulations.
2. Storing simulation and environment metadata could help to answer LANL mission critical questions.
3. Increasing FLOPS doesn’t always help answer scientific questions – only gives you more data that you can’t visualize immediately.

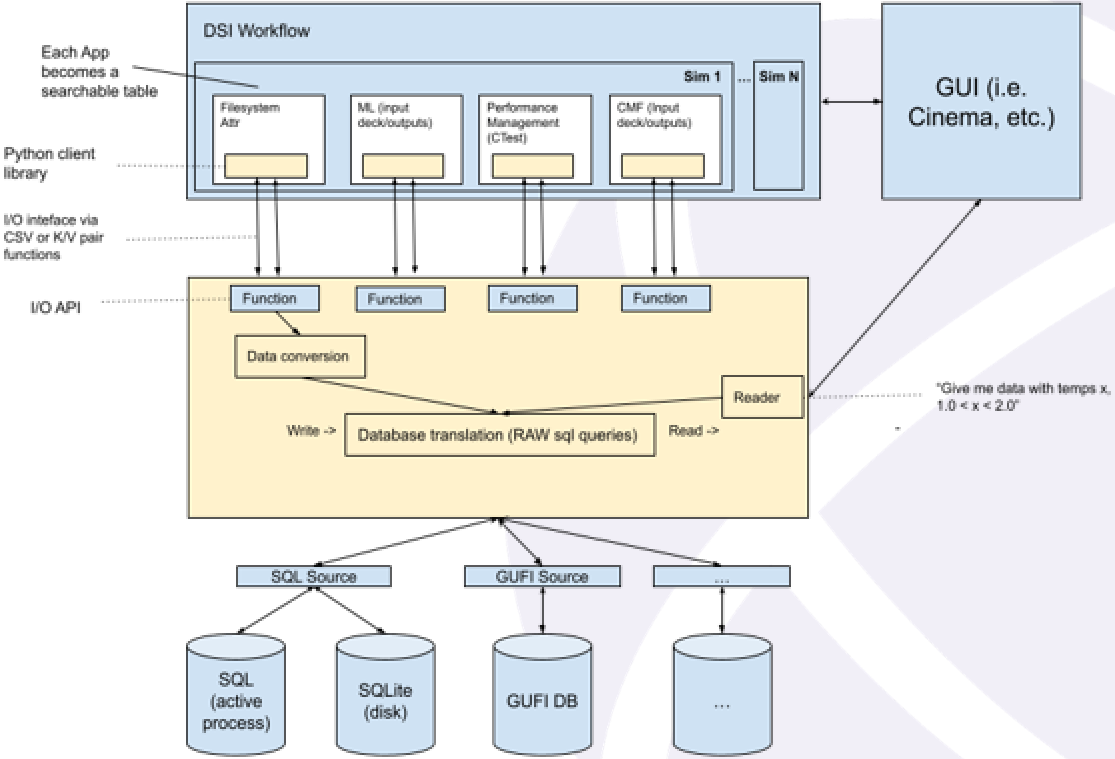


Challenges of Data Storage in a Compute-Centric World

1. POSIX file system permission for access security.
2. Interactivity and Querying is not available, which reduces opportunity for explainability.
3. Permanent data store does not exist. The current mainstream relational databases provide no extensibility to users
4. Lack of requirement & place for storing metadata in current workflows.
5. Experiments could take days, weeks, and months! Extremely computationally expensive to run simulation and reproduce the simulation itself.

Outline

- The Data Science Infrastructure (DSI) project is focused on providing data management workflows for existing LANL applications such as the Common Model Framework (CMF) via databases (SQL) through defined schema. This approach to data management uses many forms of metadata (simulation, filesystem, performance metrics) to populate a searchable database to enable capabilities for analysis, exploration and visualization.



Each App becomes a searchable table

Python client library

I/O inteface via CSV or K/V pair functions

I/O API

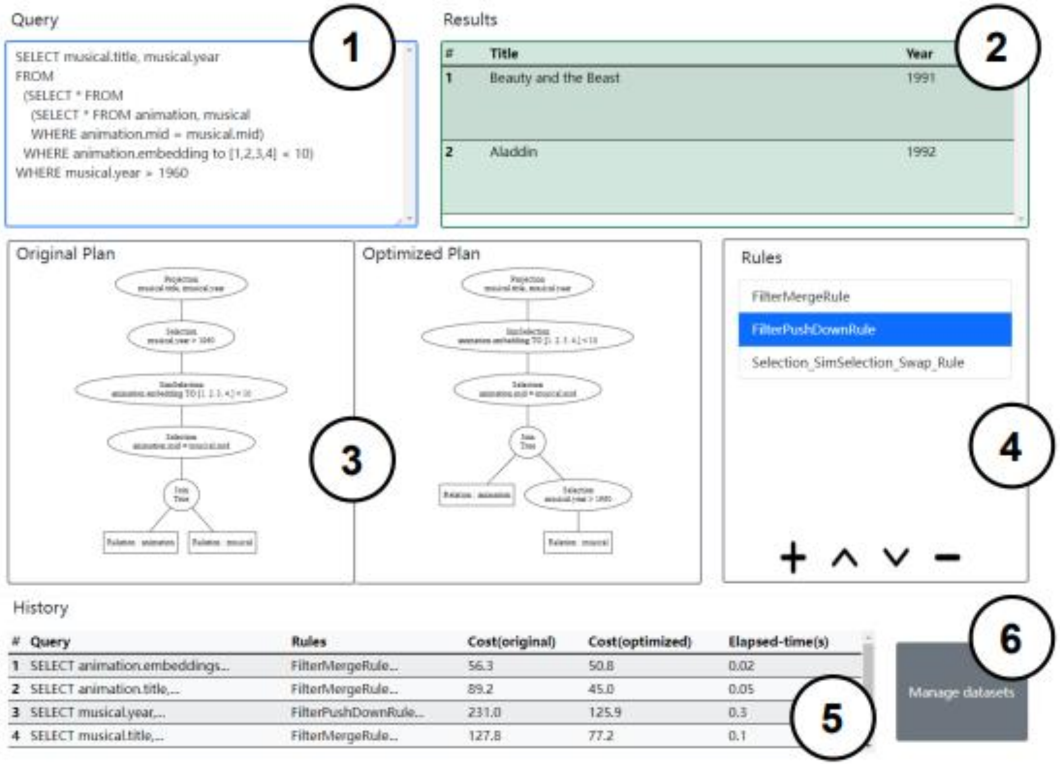
"Give me data with temps x, 1.0 < x < 2.0"

Related Works

- Extensible Database Simulator for Fast Prototyping In-Database DBSIM.

1. Implemented a system for in DB analytics of metadata.
2. GUI for SQL queries in RDBMS.

A. Difficult for LANL scientist unable to use SQL.
B. Unable to visualize analytics of metadata.


1
2
3
4
5
6
Original Plan
Optimized Plan
Projection musical.title, musical.year
Selection musical.year > 1960
SimSelection animation.embedding TO [1, 2, 3, 4] < 10
Selection animation.mid = musical.mid
Join Tree
Relation : animation
Relation : musical
+ ^ v −
Manage datasets

Query

```
SELECT musical.title, musical.year
FROM
 (SELECT * FROM
  (SELECT * FROM animation, musical
  WHERE animation.mid = musical.mid)
 WHERE animation.embedding to [1,2,3,4] < 10)
WHERE musical.year > 1960
```

Results

#	Title	Year
1	Beauty and the Beast	1991
2	Aladdin	1992

Rules

- FilterMergeRule
- FilterPushDownRule
- Selection_SimSelection_Swap_Rule

History

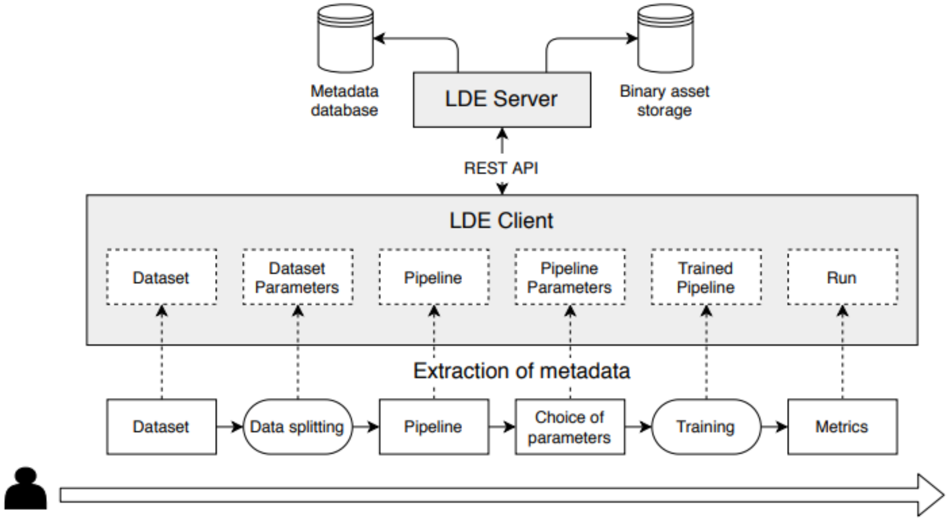
#	Query	Rules	Cost(original)	Cost(optimized)	Elapsed-time(s)
1	SELECT animation.embeddings...	FilterMergeRule...	56.3	50.8	0.02
2	SELECT animation.title,...	FilterMergeRule...	89.2	45.0	0.05
3	SELECT musical.year,...	FilterPushDownRule...	231.0	125.9	0.3
4	SELECT musical.title,...	FilterMergeRule...	127.8	77.2	0.1

Related Works Continued

- Enabling Reproducibility and Meta-learning Through a Lifelong Database of Experiments.

1. Created a permanent DB that would extract and interact with artifacts from AI experiments from datasets to pipelines.
2. Metadata extracted from experiments contained most of information to reproduce the given experiment.

A. Data captured didn’t include environment configuration and modules used.
B. No user permission restrictions.
C. No visualizations plots for statistical analysis of metadata. Unable to perform in database analytics of metadata.

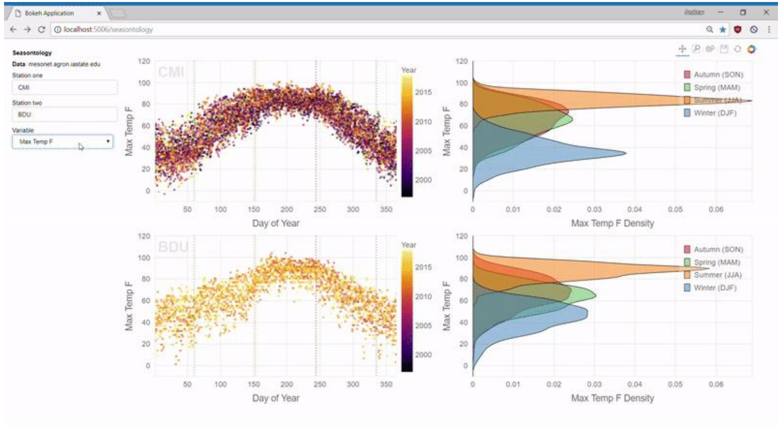


Expectations for DSI Project

- Automated Workflows and Improved metadata management for LANL mission relevant science workflows.
- Data and Experiment Reproducibility by in-database analytics and learning.
- Create a persistent database for sharing and collaboration.
- Storage and logging of parameters, code versions, metrics, environment configurations with specific modules, and output files.

Potential Front-Ends for DSI

- R&D on different GUI options for DSI such as ModelDB, Apache Superset, Trame, and ML-Flow.
- Data and Experiment Reproducibility by in-database analytics and learning.
- Create a persistent database for sharing and collaboration.


Bokeh Application

localhost:5006/seasontology

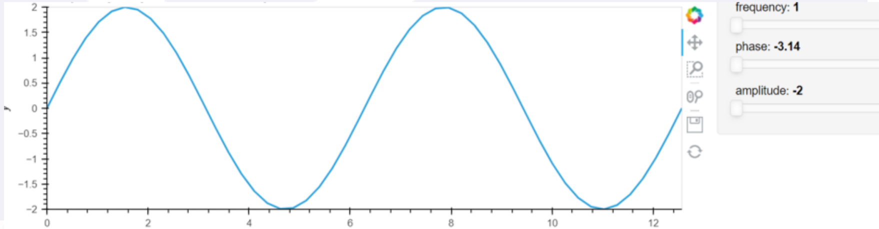
Seasontology

Data: mesonet.agron.iastate.edu

Station one: CMI

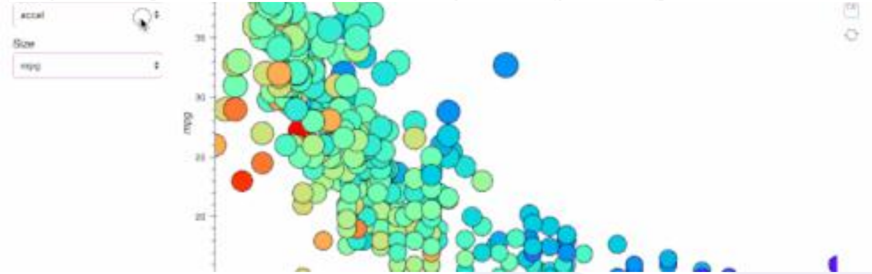
Station two: BDU

Variable: Max Temp F


frequency: 1

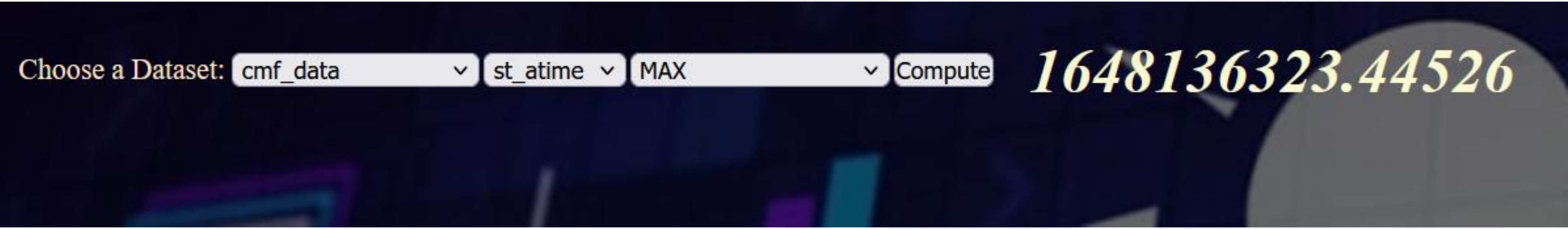
phase: -3.14

amplitude: -2



Features Implemented

- In Database analytics for comparing datasets and interactively analyze dataset statistics.
- Visualization plots for comparing metadata and sharing results.
- SQL queries implemented in backend to interact with GUI for Querying and data access API.
- Access to metadata, data via Backend APIs.

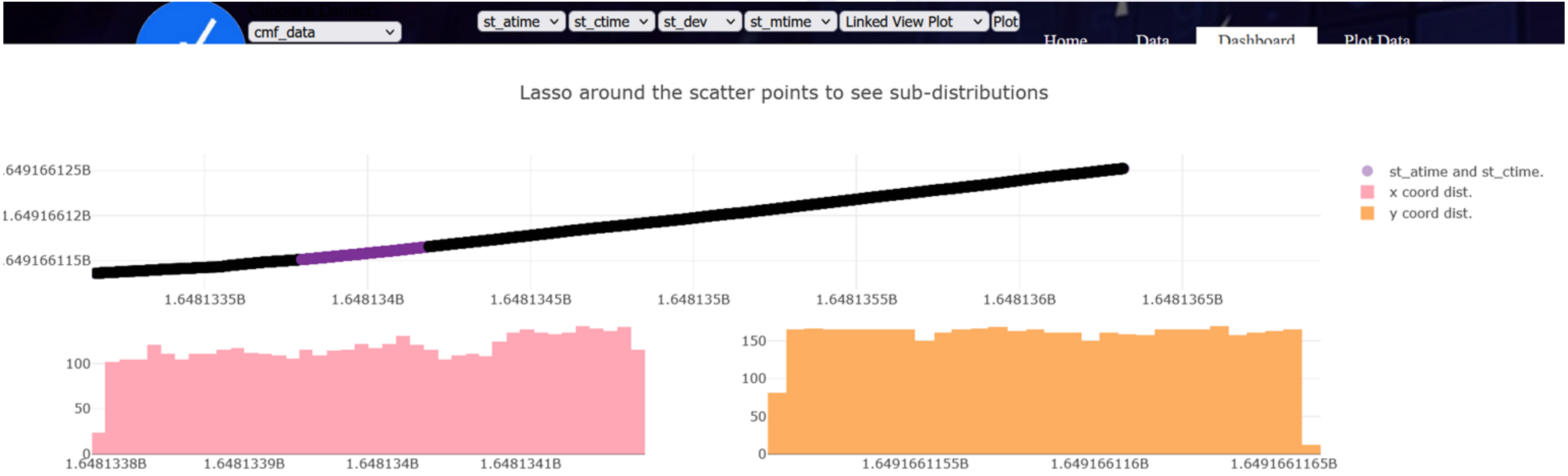


Choose a Dataset: cmf_data st_atime MAX

1648136323.44526

More Features Implemented

- Linked view of charts with point selection using the ‘lasso’ tool. Users can select points of interest on the scatter plot for an in-depth analysis and analysis of CMF data.

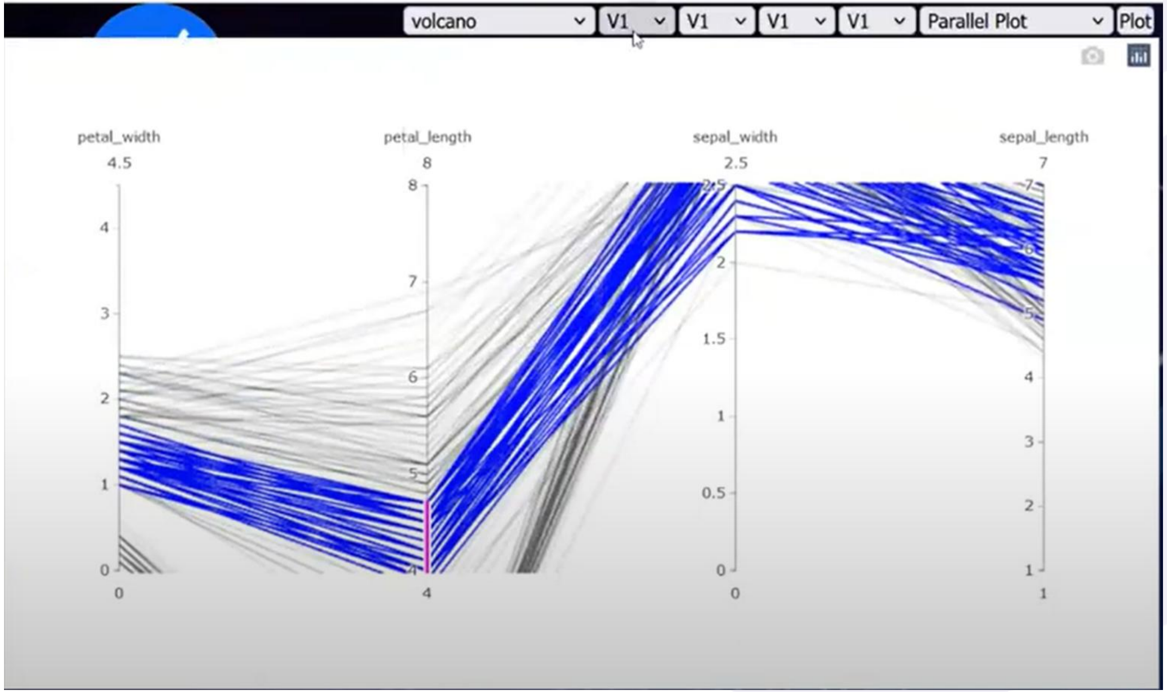


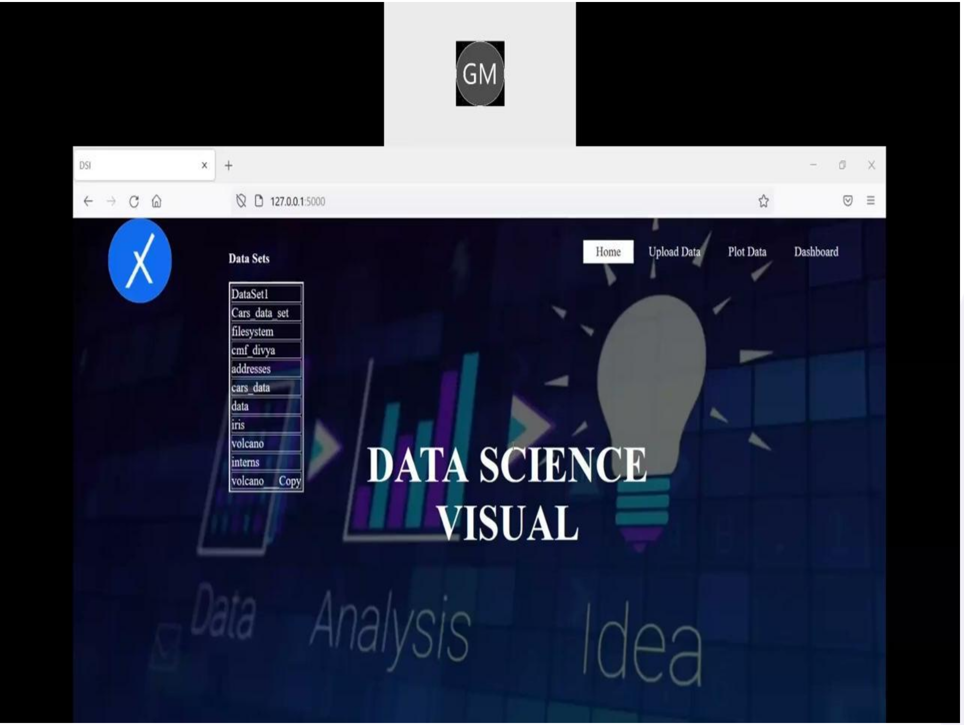
Choose a Dataset: cmf_data

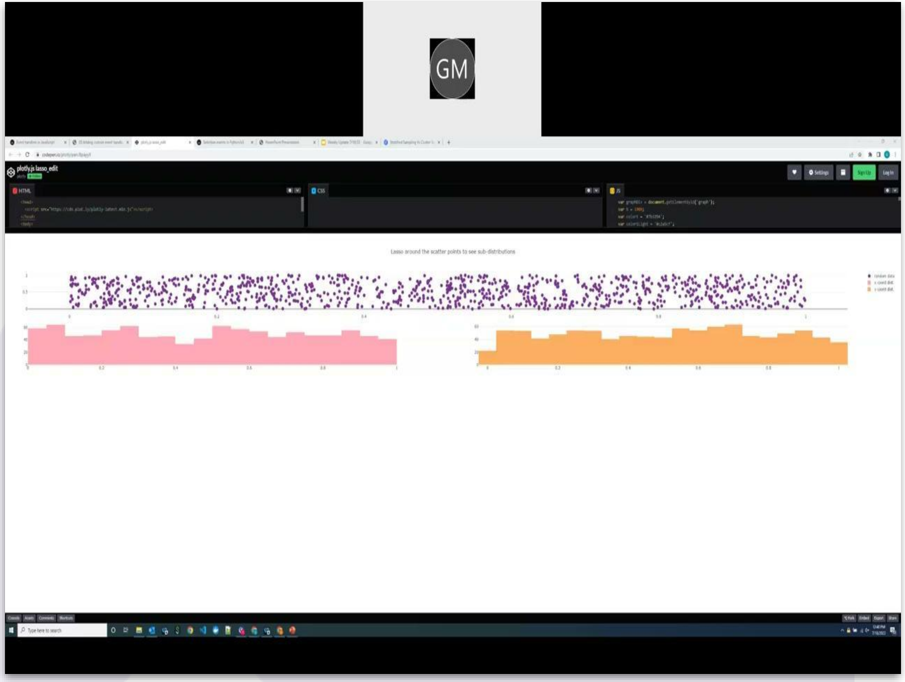
st_atime | st_ctime | st_dev | st_mtime | Linked View Plot | Plot

And More Features Implemented…

- Parallel Coordinates Plot Viewer. Selected entries from the scatter plot are highlighted in a parallel coordinates plot viewer.







```
<head>
  <script src="https://cdn.plot.ly/plotly-latest.min.js"></script>
</head>
<body>
```

```
var graphDiv = document.getElementById('graph');
var N = 1000;
var color1 = '#7b3294';
var color1Light = '#c2a5cf';
```

Lasso around the scatter points to see sub-distributions