

HPCC SYSTEMS

KSU HACKATHON



Mining the Value of Property Assessment Data



Conducting property analysis is a valuable way to understand the value of the properties in the marketplace. It is also important when making any financial decisions on whether to buy, hold, or sell.

In this hackathon, you will work with the real world property assessment data to conduct property analysis, leveraging the distributed computing environment of HPCC Systems and ECL Cloud IDE from LexisNexis Risk Solutions. A data dictionary will also be provided for a better understanding of the property assessment data.

recording_date	registry_number	sale_date	sale_price	separate_utilities	sewer	site_type	state_code	street_code
2018-10-09T00:00:00Z	009N130296	2018-10-03T00:00:00Z	438990	C	Y	A	1001	61660
2018-12-11T00:00:00Z	009N130298	2018-11-27T00:00:00Z	350000	C	Y	A	1001	61660
2018-11-28T00:00:00Z	009N130299	2018-10-17T00:00:00Z	448000	C	Y	A	1001	61660

Partial Sample Dataset

Project Evaluation:

- **You will be rated on how you break down the problem, how you understand the data, how you shape the data for analysis and then the steps you take to analyze it.**
- HPCC Systems Mentor team will base on below criteria to measure the project performance:
 1. Please record each step clearly in the source code in ECL Cloud IDE
 2. Report what's discovered in each step in the final presentation.
 3. Play each step as a demo in ECL Cloud IDE in the final presentation.

For example, if you execute a data profiling, please record this step clearly in the source code in ECL Cloud IDE and report what's discovered in the final presentation.

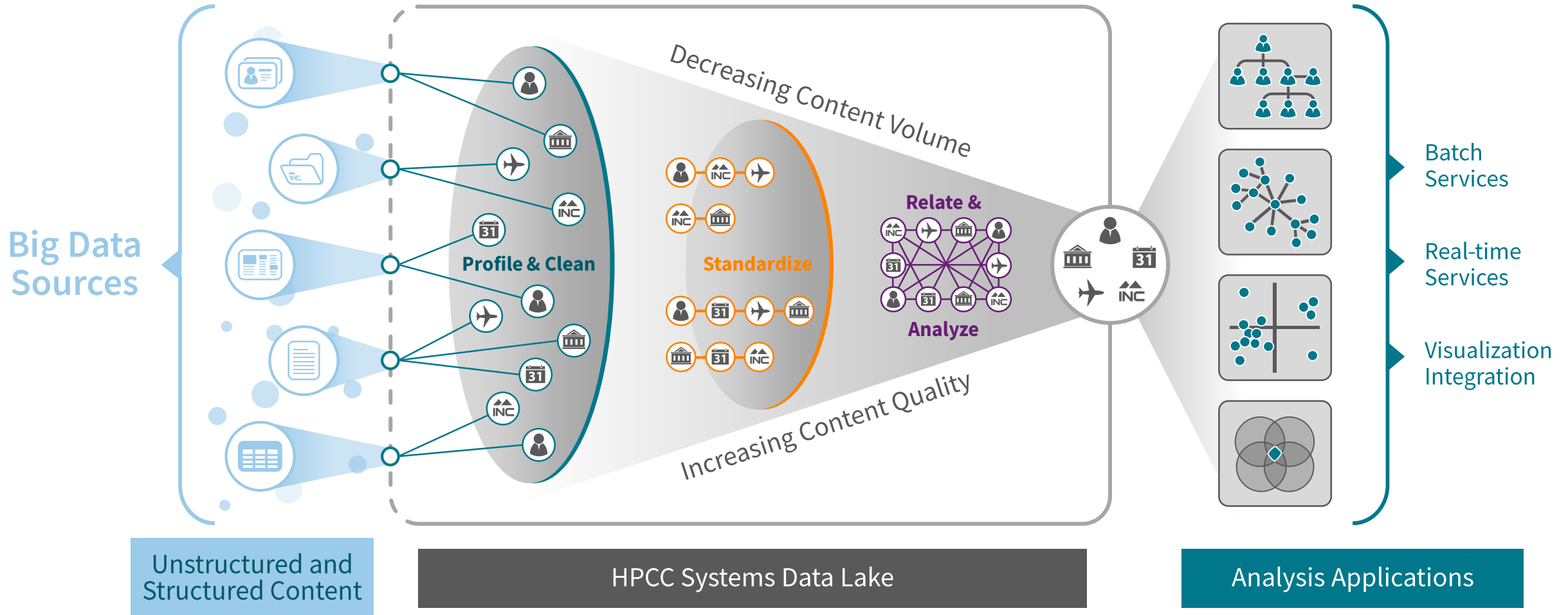
Process is more important than the result





What you will need in this hackathon?

HPCC Systems (Small to Big Data) ETL

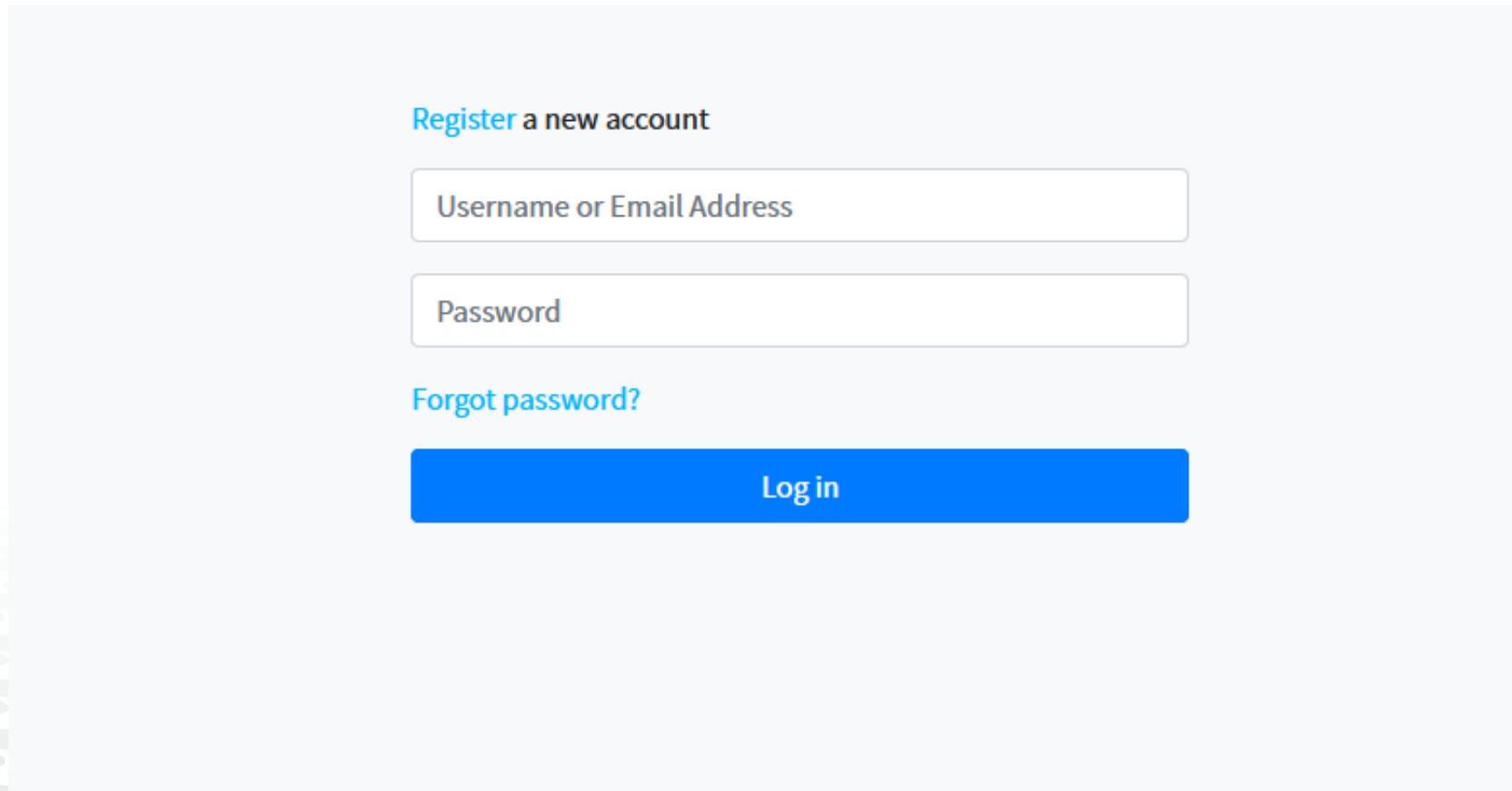


Machine Learning on HPC Systems Platform



ECL Cloud IDE

<https://ide.hpccsystems.com/>

A screenshot of the ECL Cloud IDE login page. The page has a light gray background. At the top, there is a link "Register a new account" in blue text. Below this are two input fields: "Username or Email Address" and "Password". Below the password field is a link "Forgot password?" in blue text. At the bottom is a blue button with the text "Log in" in white.

[Register a new account](#)

[Forgot password?](#)
[Log in](#)

ECL Cloud IDE

The screenshot displays the ECL Cloud IDE workspace. The top navigation bar includes the 'ECL IDE' logo, a workspace selector 'ksu_workshop', and buttons for 'NEW +' and 'DELETE'. On the right, there are links for 'Help', a user profile 'lily', and a 'Logout' button.

The left sidebar is divided into two sections: 'DATASETS' and 'SCRIPTS'. Under 'DATASETS', there are 'NySampleinput' and 'Sample10000'. Under 'SCRIPTS', there is a folder 'HPCC-ECL-Training' containing sub-folders 'NYTaxiTrip' and 'Hackathon_Seed_Project'. The 'NYTaxiTrip' folder contains scripts 'A_Data_Ingestion', 'B_Data_Validation', 'C_Data_Profiling', 'D_Data_Enhancement', 'E_LinearRegression', and 'F_LogisticRegression'. The 'Hackathon_Seed_Project' folder contains 'A_Data_Ingestion', 'B_Data_Validation', 'C_Data_Profiling', and 'D_Data_Enhancement'.

The main area is titled 'OUTPUTS' and shows 'Result 1' with a 'Data Patterns' tab. It displays two data pattern cards:

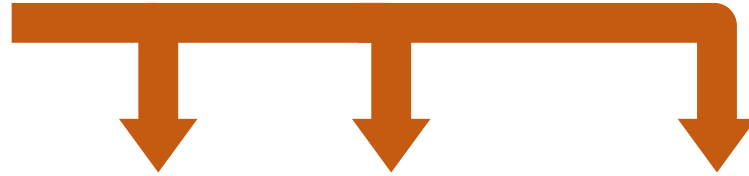
Field	Min Length	Avg Length	Max Length	Cardinality	Popular Patterns
number_of_rooms	1	1	2	6 (39%)	9 (99%)
assessment_date	10	10	10	1949-01-01 (100%)	9999-99-99 (100%)

Below the data patterns, there is a 'RUN' button and a code editor showing the following ECL script:

```
1 IMPORT STD;
2 IMPORT DataPatterns;
3
4 //Step 1 : read in the raw data
5 Layout := RECORD
6   STRING number_of_rooms;
7   STRING assessment_date;
8   STRING beginning_point;
9   STRING book_and_page;
10  STRING building_code;
11  STRING building_code_description;
12  STRING category_code;
13  STRING category_code_description;
14  STRING census_tract;
15  STRING central_air;
16  STRING cross_reference;
17  STRING date_exterior_condition;
18  STRING depth;
```

ECL Cloud IDE WorkSpace

HPCC Systems KSU Hackathon 2019



<https://hpccsystems.com/hpccsummit2019>



CONNECT



SOLVE

PRESENT

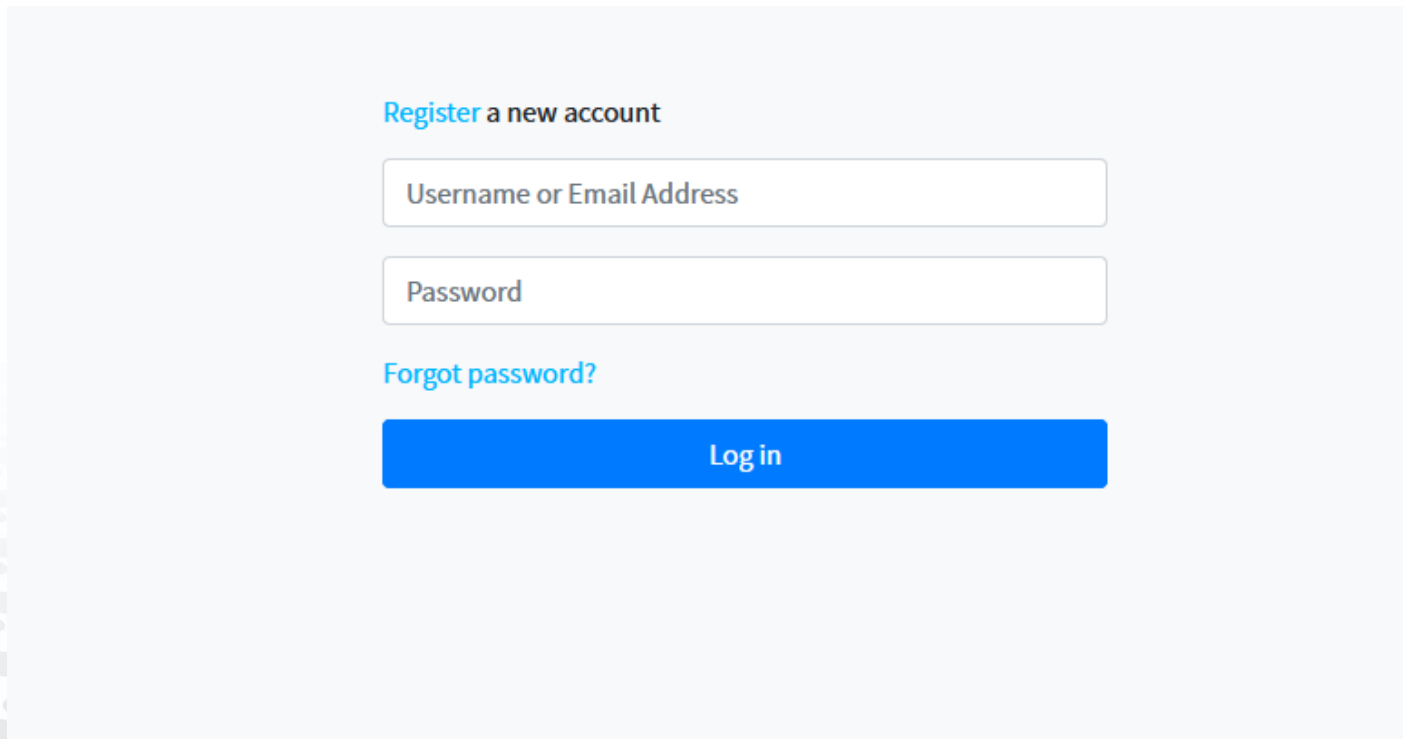


ECL IDE

Hackathon Tips

1. Register to become a member of **ECL Cloud IDE** on KSU Campus

<https://ide.hpccsystems.com/auth/login>

A screenshot of the ECL Cloud IDE login page. The page has a light gray background. At the top, there is a link "Register a new account" in blue text. Below this are two input fields: "Username or Email Address" and "Password". Below the password field is a link "Forgot password?" in blue text. At the bottom is a blue button with the text "Log in" in white.

[Register a new account](#)

[Forgot password?](#)

Log in

Hackathon Tips – cont.

2. Copy the KSU_Hackathon WorkSpace into your ECL Cloud IDE by open below link in the browser:

<https://ide.hpccsystems.com:/workspaces/share/758976f6-49e2-4f18-9395-9fb86850e6b6>

Hackathon Tips – cont.

3. Try NYTaxiTrip examples in HPCC-ECL-Training folder to rewind the workshop

The screenshot displays the HPCC ECL IDE interface. On the left sidebar, under 'DATASETS', there are 'NySampleinput' and 'Sample10000'. Under 'SCRIPTS', there is a folder 'HPCC-ECL-Training' containing 'NYTaxiTrip', 'A_Data_Ingestion', 'B_Data_Validation', 'C_Data_Profiling', 'D_Data_Enhancement', 'E_LinearRegression', 'F_LogisticRegression', and 'Hackathon_Seed_Project'. The main area shows the 'OUTPUTS' tab with a table of data. The table has columns: 'id', 'month_of_year', 'day_of_week', 'precipintensity', and 'trip_counts'. The first four rows of data are visible. Below the table is a 'RUN' button and a script editor with the following code:

```
1  IMPORT ML_Core;  
2  IMPORT ML_Core.Types;  
3  IMPORT NYTaxiTrip.D_Data_Enhancement;  
4  IMPORT LinearRegression AS LROLS;  
5
```

Hackathon Tips – cont.

- **NOTE:**

The name convention to refer the file uploaded to your workspace is

'~USERNMAE::WORKSPACENAME::RAWFILENAME'

Example:

If your username is Mike, you created a workspace 'HPCCSystems' and uploaded the file 'test.csv' to the workspace.

To use the dataset in ECL Cloud IDE, the directory to refer the file should be

'~Mike::HPCCSystems::test.csv'.

To ingest the dataset in ECL Cloud IDE, you can use DATASET function, such as:

raw := DATASET('~Mike::HPCCSystems::test.csv', Layout, CSV(HEADING(1)));

Hackathon Tips – cont.

4. Try the Hackathon_Seed_Project in ECL Cloud IDE.

This project shows a few sample steps that help you familiar with the Property Assessment sample dataset -- Sample10000. It's uploaded to the shared WorkSpace KSU_Workshop.

The screenshot displays the ECL IDE interface for the 'ksu_workshop' workspace. The left sidebar contains a 'DATASETS' section with 'NySampleinput' and 'Sample10000', and a 'SCRIPTS' section with 'HPCC-ECL-Training', 'NYTaxiTrip', 'Hackathon_Seed_Project', 'A_Data_Ingestion', 'B_Data_Validation', 'C_Data_Profiling', and 'D_Data_Enhancement'. The main area shows 'OUTPUTS Result 1' with a table of 25 entries. The table has columns: 'number_of_rooms', 'assessment_date', 'beginning_point', 'book_and_page', and 'building_code'. The first three rows are visible, each starting with a '0' in the 'number_of_rooms' column. Below the table is a 'RUN' button and a code editor showing a script to read raw data from the HPCC Systems cluster.

number_of_rooms ↕	assessment_date ↕	beginning_point ↕	book_and_page ↕	building_code ↕
0		32'6" S TASKER ST	0602048	U50
0		48'6" S TASKER ST	3400021	O50
0		64'6" S TASKER ST	1416545	O50

```
1 //Read raw data from HPCC Systems cluster
2
3 Layout := RECORD
4   STRING number_of_rooms;
5   STRING assessment_date;
6   STRING beginning_point;
7   STRING book_and_page;
```

Hackathon Tips – cont.

- **Note:**

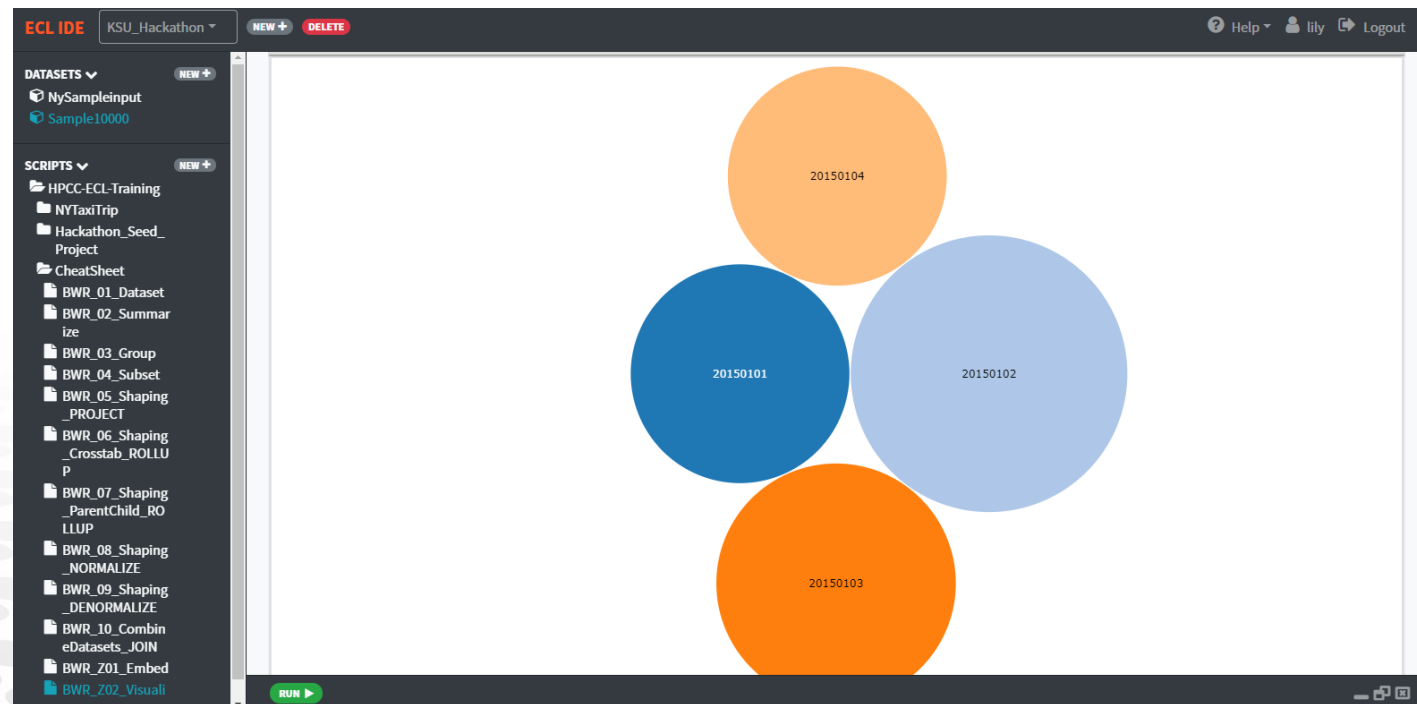
- The sample dataset Sample10000 only includes 10,000 records of the original Property Assessment dataset. Your solution should apply to the original Property Assessment Dataset as final result.
- The original Property Assessment dataset can be accessed as below in A_Data_Ingestions file:

```
propertyDS := DATASET('~ksu::hackathon::opa_properties_public.csv', Layout,  
                      CSV(HEADING(1)));
```


Hackathon Tips – cont.

5. Try the CheatSheet in ECL Cloud IDE.

This project is a great resource to quick start ECL programming. It includes most frequently used ECL functions such as DATASET, PROJECT, TABLE, ROLLUP, NORMALIZE and also includes functions to visualize your data in HPCC Systems.



Hackathon Tips – cont.

6. Understand Property Assessment Data via Data Dictionary:

https://github.com/lilyclemons/KSU_Hackathon2019

Hackathon Tips – cont.

7. Other HPCC Systems KSU hackathon related ECL Code examples and slides are available at Git repository KSU_Hackathon2019 :

https://github.com/lilyclemson/KSU_Hackathon2019

Hackathon Tips – cont.

8. Welcome to join our **#hpcc** Channel to interact with mentors and ask questions via below slack space or click on below link:

#ksuccsehackathon: ksuccsehackathon.slack.com

https://join.slack.com/t/ksuccsehackathon/shared_invite/enQtNzY0MjY0NjQyNzc0LTUzYzYzMTBhYmIzMjFmNjlkYTYwZDRmMGI3MzFkOGE1ZTQ2MDIzYWM3MGEzZjJkZTMwNjJjNmM5ZDNIOWFhYzQ

Helpful Links

- Introduction of HPCC System
<https://hpccsystems.com/about>
- ECL CheatSheet:
<https://github.com/hpcc-systems/HPCC-ECL-Training/tree/master/CheatSheet>
- Introduction of HPCC Systems Machine Learning Library
<https://hpccsystems.com/download/free-modules/machine-learning-library>
- ECL Machine Learning Examples:
<https://github.com/hpcc-systems/HPCC-ECL-Training/tree/master/NYTaxiTrip>
<https://github.com/lilyclemson/HPCC-ECL-Training/tree/master/StockTrade>
- Other Documentations
<https://hpccsystems.com/training/documentation>
- Opportunity to attend HPCC-Systems-Summit-2019
<https://hpccsystems.com/community/events/hpcc-systems-summit-2019>

