

SAMSUNG

Samsung Innovation Campus

| **Khoá học Big Data**

Together for Tomorrow!
Enabling People

Education for Future Generations

Chương 8

Trực quan hóa dữ liệu

Khoá học Big Data

Mô tả chương

📌 Mục tiêu:

- ✓ Trong chương này, chúng ta sẽ tìm hiểu về cách sử dụng trực quan hóa dữ liệu của mình. Có nhiều công cụ nguồn mở để lựa chọn nhưng vì chúng tôi đang sử dụng nền tảng Hadoop nên chúng tôi sẽ giới thiệu **HUE** (Trải nghiệm người dùng Hadoop) và các khả năng trực quan hóa mà nó cung cấp
- ✓ Sau đó, chúng ta sẽ chuyển sang thảo luận tổng quát hơn nhiều về Trực quan hóa bằng các sản phẩm thương mại. So với các mã nguồn mở, các sản phẩm thương mại có nhiều tính năng phong phú hơn.
 - Mặc dù có nhiều sản phẩm thương mại tốt, nhưng chúng tôi sẽ tập trung vào MS Power BI trong chương này. MS Power BI là công cụ trực quan được xếp hạng cao nhất trong báo cáo Gartner năm 2021 và có phiên bản miễn phí để chúng ta có thể dễ dàng trải nghiệm thực tế.

📌 Nội dung:

1. Công cụ trực quan hóa mã nguồn mở
2. Công cụ trực quan hóa thương mại phổ biến

Bài 1

Công cụ trực quan hóa mã nguồn mở

Trực quan hoá dữ liệu

Bài 1

Công cụ trực quan hóa mã nguồn mở

| 1.1. Khái niệm cơ bản về trực quan hóa dữ liệu

| 1.2. Giới thiệu về Apache Hue và Jupyter

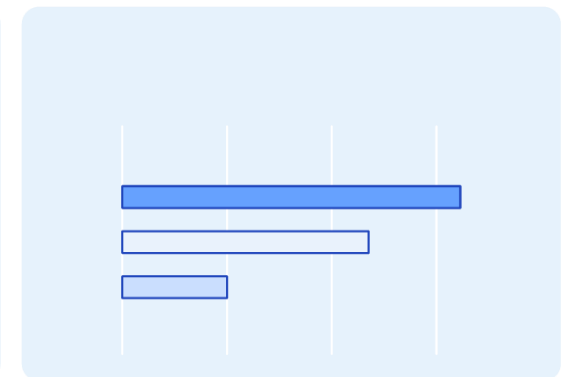
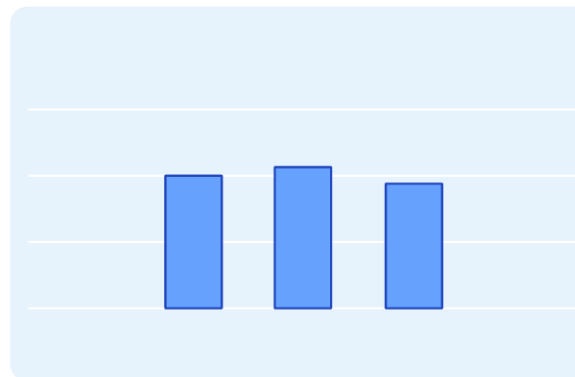
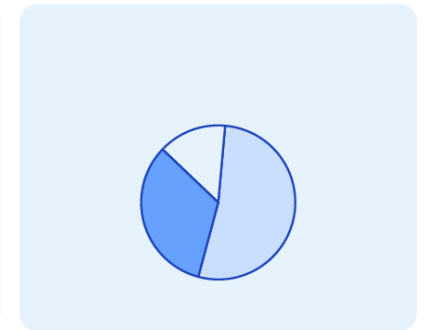
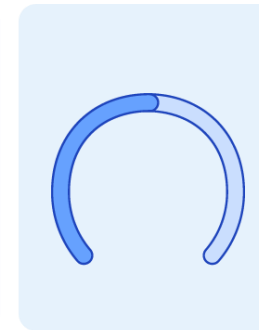
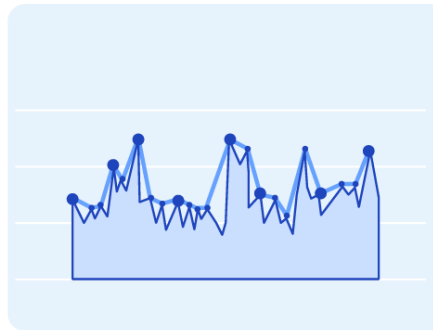
Tại sao cần trực quan hóa?

- I Dữ liệu có một câu chuyện để kể
 - ▶ Trực quan hóa hiệu quả sẽ kể câu chuyện đó
- I Trực quan hóa dữ liệu là một hình thức nghệ thuật giúp thu hút sự chú ý của chúng tôi và tập trung vào đúng thông điệp
- I Trực quan hóa tốt giúp chúng tôi nhanh chóng nhìn thấy xu hướng và hiểu dữ liệu cơ bản
- I Trực quan hóa loại bỏ mọi rào cản kỹ thuật hoặc toán học để hiểu dữ liệu
 - ▶ Một bảng tính lớn với các hàm và công thức liên quan đến nhau sẽ không truyền tải được gì cho những người không chuyên về kỹ thuật.



Lợi ích của trực quan hóa hiệu quả

- I Khán giả có thể xem và tiếp thu thông tin dễ dàng hơn nhiều
- I Xác định các mô hình và xu hướng mới nổi
 - ▶ Giúp liên kết chiến lược hoạt động của tổ chức với kết quả kinh doanh
 - ▶ Dự đoán cơ hội bán hàng và doanh thu
- I Giúp các tổ chức xác định các vấn đề và thiếu sót và thực hiện các điều chỉnh nhanh chóng giúp ích cho lợi nhuận của họ



Trực quan hóa chính xác

- I Một bài thuyết trình và hình ảnh bên trong nó có đối tượng mục tiêu cụ thể
 - ▶ Hình ảnh hóa phải được thiết kế có tính đến đối tượng và nhu cầu của họ
 - ▶ Tiếp thị và bán hàng
 - ▶ Kế hoạch sản phẩm
 - ▶ Tài chính và kế toán
- I Chọn các kỹ thuật trực quan hóa phù hợp cho loại dữ liệu cơ bản
 - ▶ Bàn để hồ sơ
 - ▶ Biểu đồ đường để theo dõi các thay đổi
 - ▶ Biểu đồ đường để so sánh số lượng



Công thức để làm tốt nhiệm vụ trực quan hóa

I Dữ liệu tốt

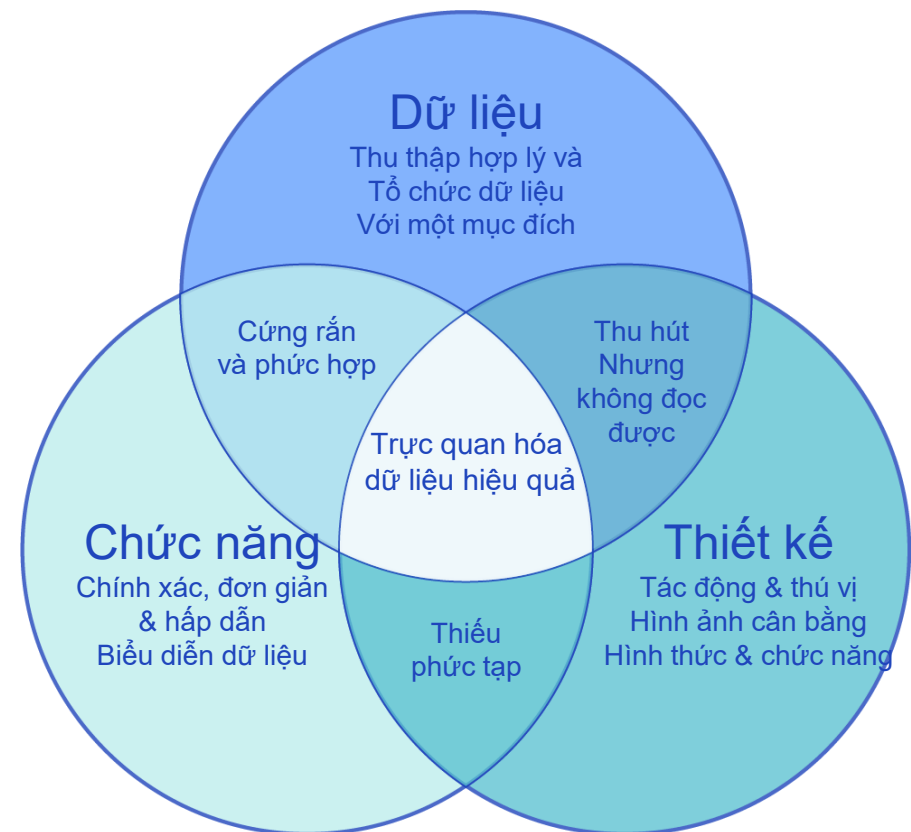
- ▶ Rác vào → Rác ra.
- ▶ Dữ liệu từ nhiều nguồn, được tích hợp tốt, là thành phần ban đầu để có một hình ảnh trực quan tốt

I Chức năng tốt

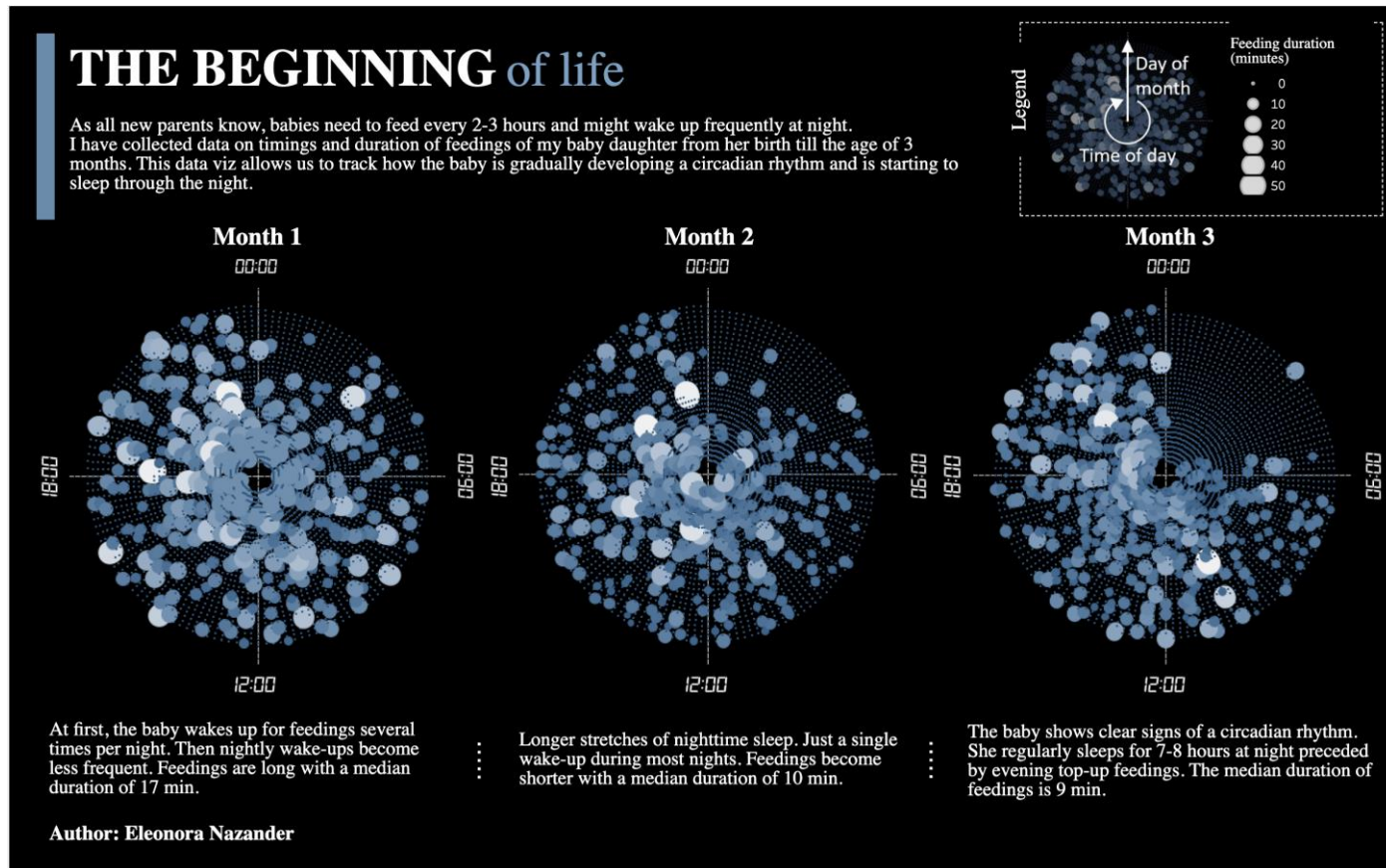
- ▶ Biểu diễn chức năng của dữ liệu giúp dễ nhìn và dễ hiểu hơn
- ▶ Trộn và kết hợp, cắt và thái - một đầu bếp giỏi biết cách lấy nguyên liệu và biến nó thành một món ăn đặc biệt

I Thiết kế tốt

- ▶ Hình ảnh thú vị khiến khán giả quan tâm đến dữ liệu
- ▶ Nghệ thuật và hình thức - món ăn không hoàn chỉnh nếu không có phần trình bày hoàn thiện



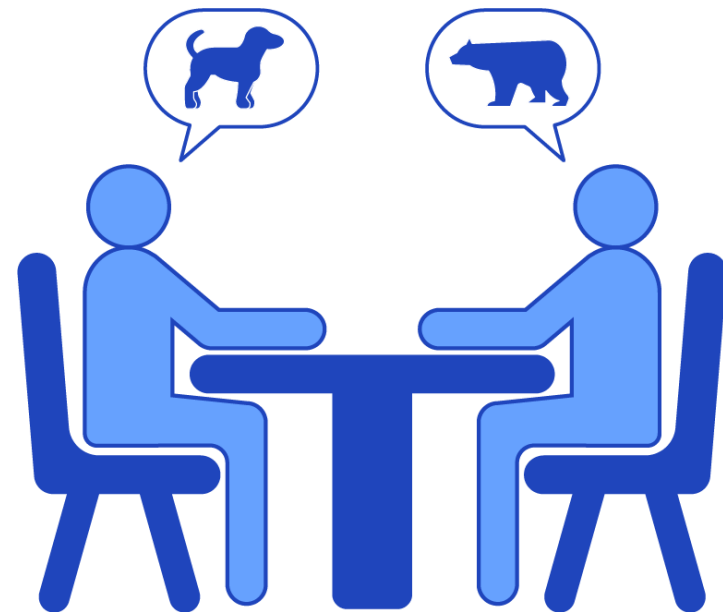
Ví dụ về chức năng và hình thức



Tác giả: Eleonora Nazander
Nguồn: Tableau Public Gallery

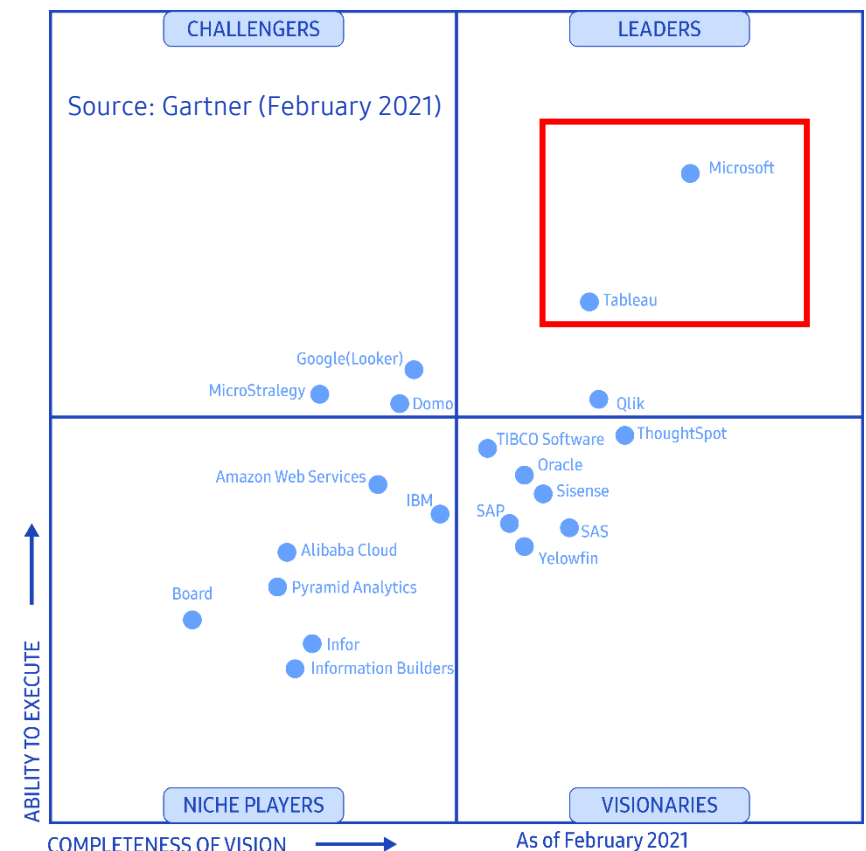
Cuối cùng nhưng không kém phần quan trọng - Trên thực tế, ĐẦU TIÊN!!!

- I Dữ liệu, thiết kế, chức năng đều là những thành phần chính để có một bản trình bày tốt
- I Tuy nhiên, một hình ảnh trực quan tốt bắt đầu trước khi một biểu đồ được tạo
- I Nó bắt đầu với cuộc phỏng vấn khách hàng
 - ▶ Câu hỏi mà khách hàng đang cố gắng trả lời là gì?
 - ▶ Dữ liệu có thể được thu thập và sắp xếp ở đâu và như thế nào để trả lời câu hỏi
 - ▶ Đảm bảo rằng bạn và khách hàng hiểu đầy đủ câu hỏi và kết quả mong đợi



Gartner Magic Quadrant

- I Gartner's Magic Quadrant là một tiêu chuẩn ngành để các chuyên gia CNTT so sánh và hiểu rõ hơn về các sản phẩm cạnh tranh khác nhau hiện có trên thị trường
- I Trực X là sự hoàn chỉnh của tầm nhìn
 - ▶ Trực này cho biết sản phẩm nào có các tính năng cơ bản cũng như tầm nhìn của sản phẩm khiến sản phẩm trở nên nổi bật giữa đám đông
- I Trực Y là khả năng thực thi
 - ▶ Trực này cho thấy mỗi sản phẩm đã có thể thực hiện tốt như thế nào với tầm nhìn của họ với việc phân phối thực tế thành một sản phẩm dễ sử dụng và được cân nhắc kỹ lưỡng
- I Góc phần tư trên cùng bên phải đại diện cho các sản phẩm có tầm nhìn đầy đủ và cũng có thể thực hiện theo tầm nhìn đó



Bài 1

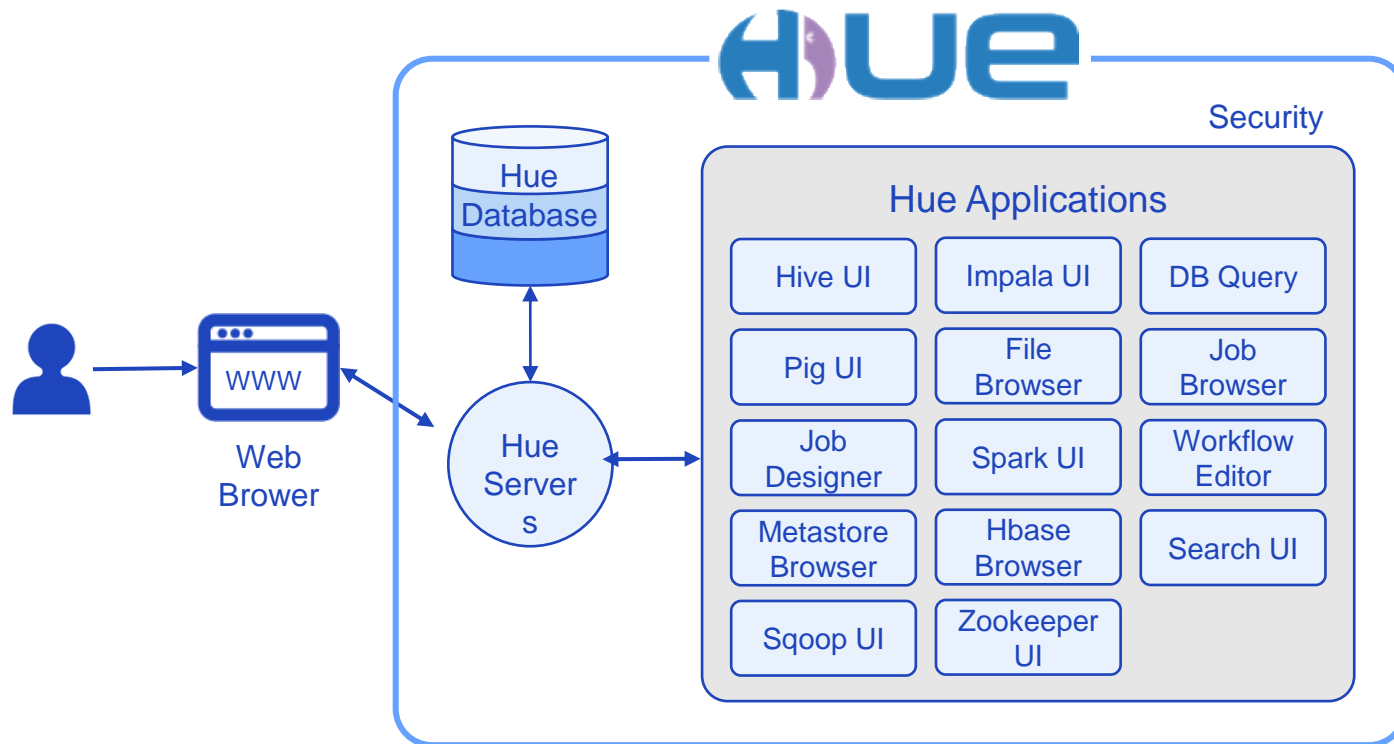
Công cụ trực quan hóa mã nguồn mở

| 1.1. Khái niệm cơ bản về trực quan hóa dữ liệu

| 1.2. Giới thiệu về Apache Hue và Jupyter

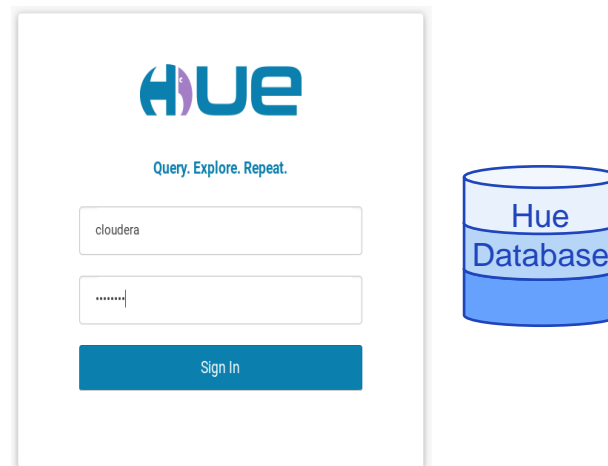
HUE là gì?

- Cung cấp giao diện web để tương tác với cụm Hadoop



Truy cập HUE

- | SuperUser (Người dùng đầu tiên đăng nhập)
- | Truy cập **Hue** từ trình Duyệt
 - ▶ `http://<hue_server>:8888` (Non-secure , w/o Hue balancer)
- | Người dùng lần đầu tiên đăng nhập vào **Hue** có đặc quyền Superuser
 - ▶ Superuser có quyền tạo và quản lý người dùng và nhóm, đồng thời chỉ định superuser



Tại sao chọn HUE?

I Các chức năng chính của Hue

Truy vấn

Hive, Impala query editor, Pig editor, ...

Tìm kiếm

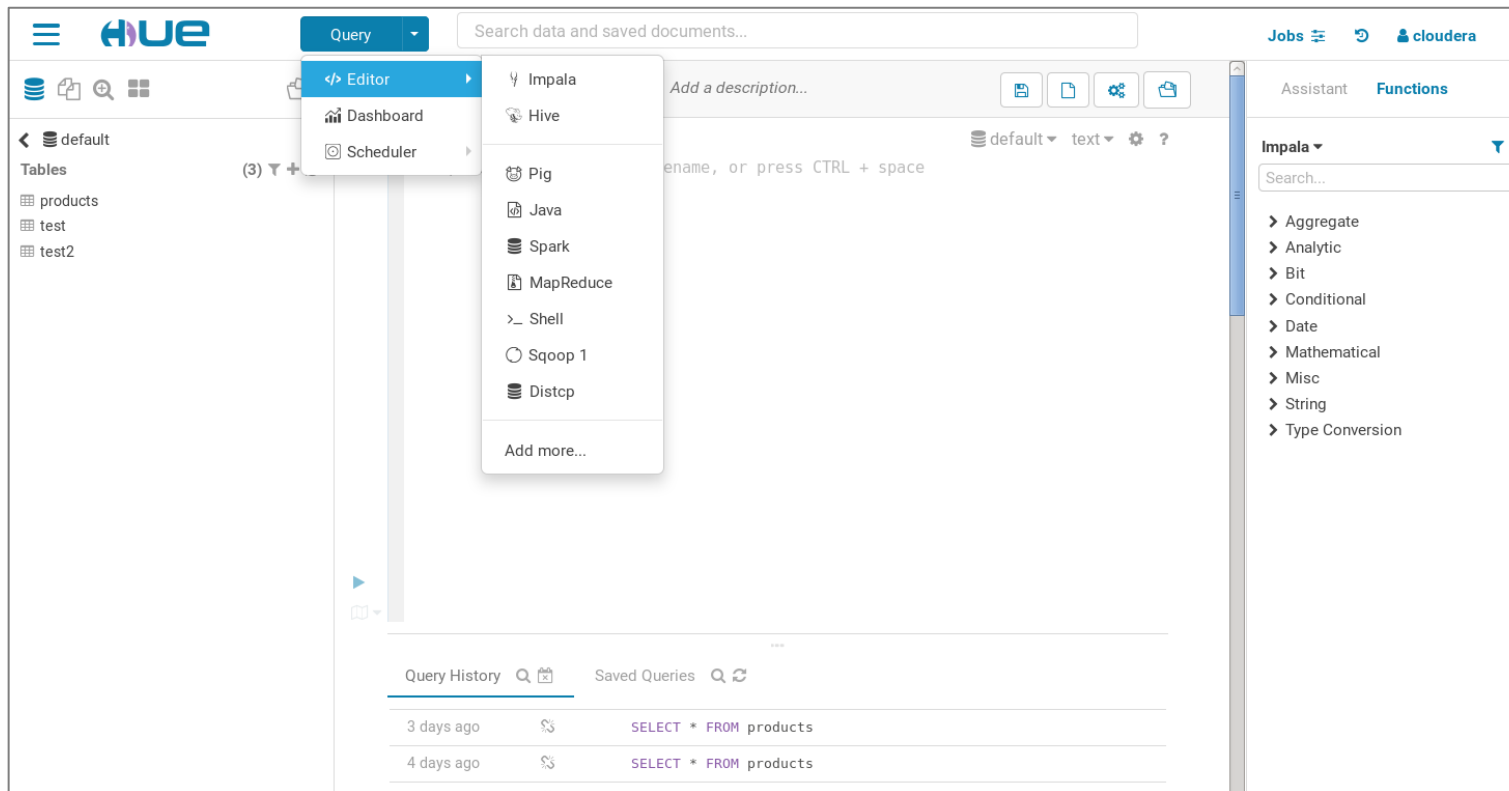
File browser, job browser, ...

Lên lịch

Oozie workflow Web UI, ...

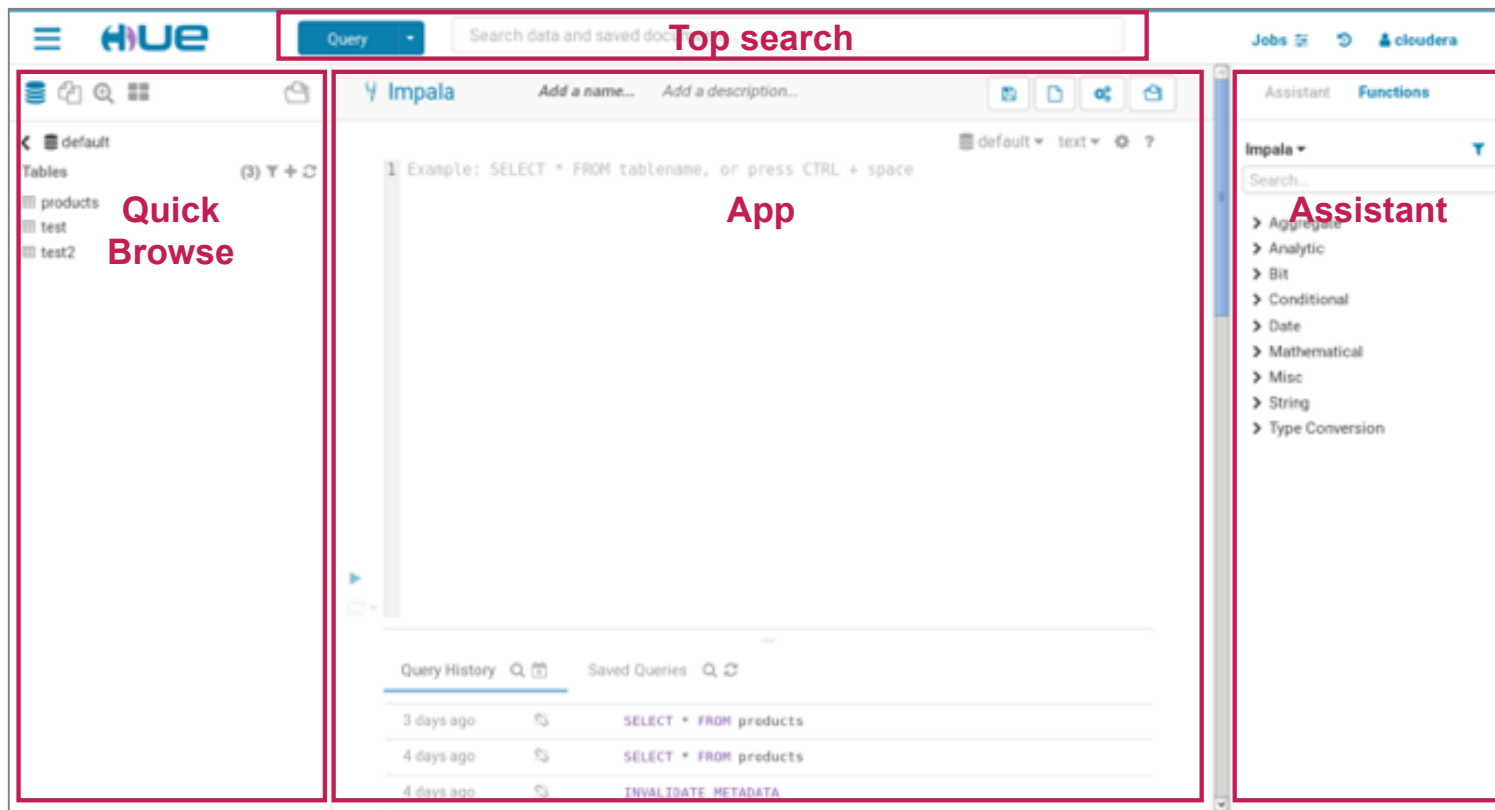
Hue : Trình chỉnh sửa truy vấn

I Cung cấp Web Truy Vấn UI cho các dịch vụ Hadoop khác nhau



Hue : Truy vấn (1/6)

I Khu vực chính



Hue : Truy vấn (2/6)

I Trình chỉnh sửa truy vấn cho Hive / Impala

The screenshot displays the Apache Hue query editor interface. A red box highlights the main query area, labeled "Space to write Query". The query is written in Hive SQL and includes comments. Annotations with blue arrows point to specific features:

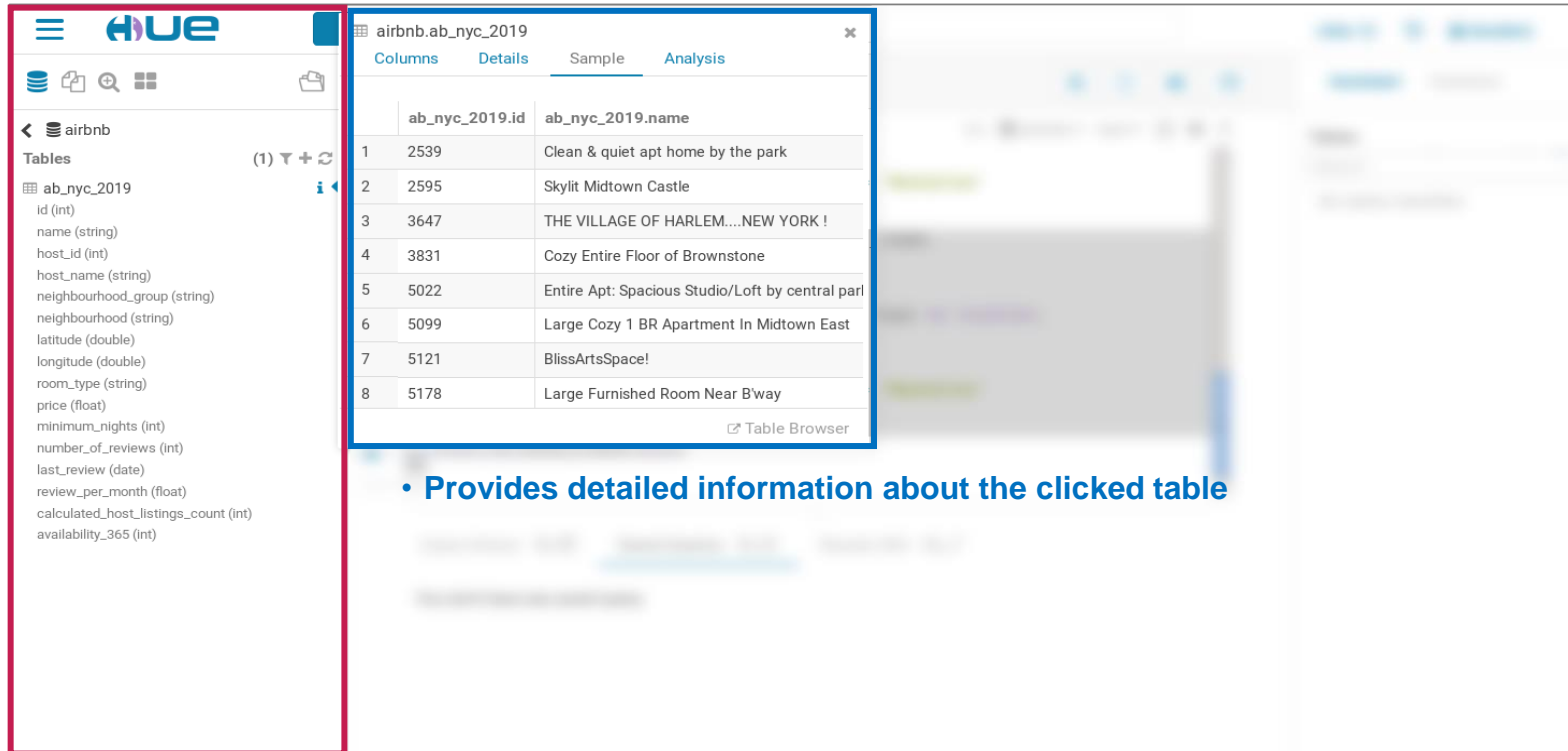
- Provides Query Save and Load functions**: Points to the "Add a name..." and "Add a description..." buttons at the top of the query editor.
- DB in use**: Points to the "airbnb" database selector in the top right corner.
- Executable only for the specified block**: Points to the "Run" button (a blue triangle) on the left side of the query editor.

```
30 select(neighbourhood, , neighbourhood_group, as location,
31 room_type,
32 price as `price($)`
33 from ab_nyc_2019
34 where price < 150.0 and neighbourhood_group = 'Manhattan'
35 order by `price($)`,location desc;
36
37 select location, room_type, count(*) as avail_rooms
38 from (
39 select name as room_descriptions,
40 host_name,
41 concat(neighbourhood, ', ', neighbourhood_group) as location,
42 room_type,
43 price as `price($)`
44 from ab_nyc_2019
45 where price < 150.0 and neighbourhood_group = 'Manhattan'
46 ) t1
47 group by location, room_type
48 order by avail_rooms desc;
```

Hue : Truy vấn (3/6)

I Trình chỉnh sửa truy vấn cho Hive / Impala

DB / Table information



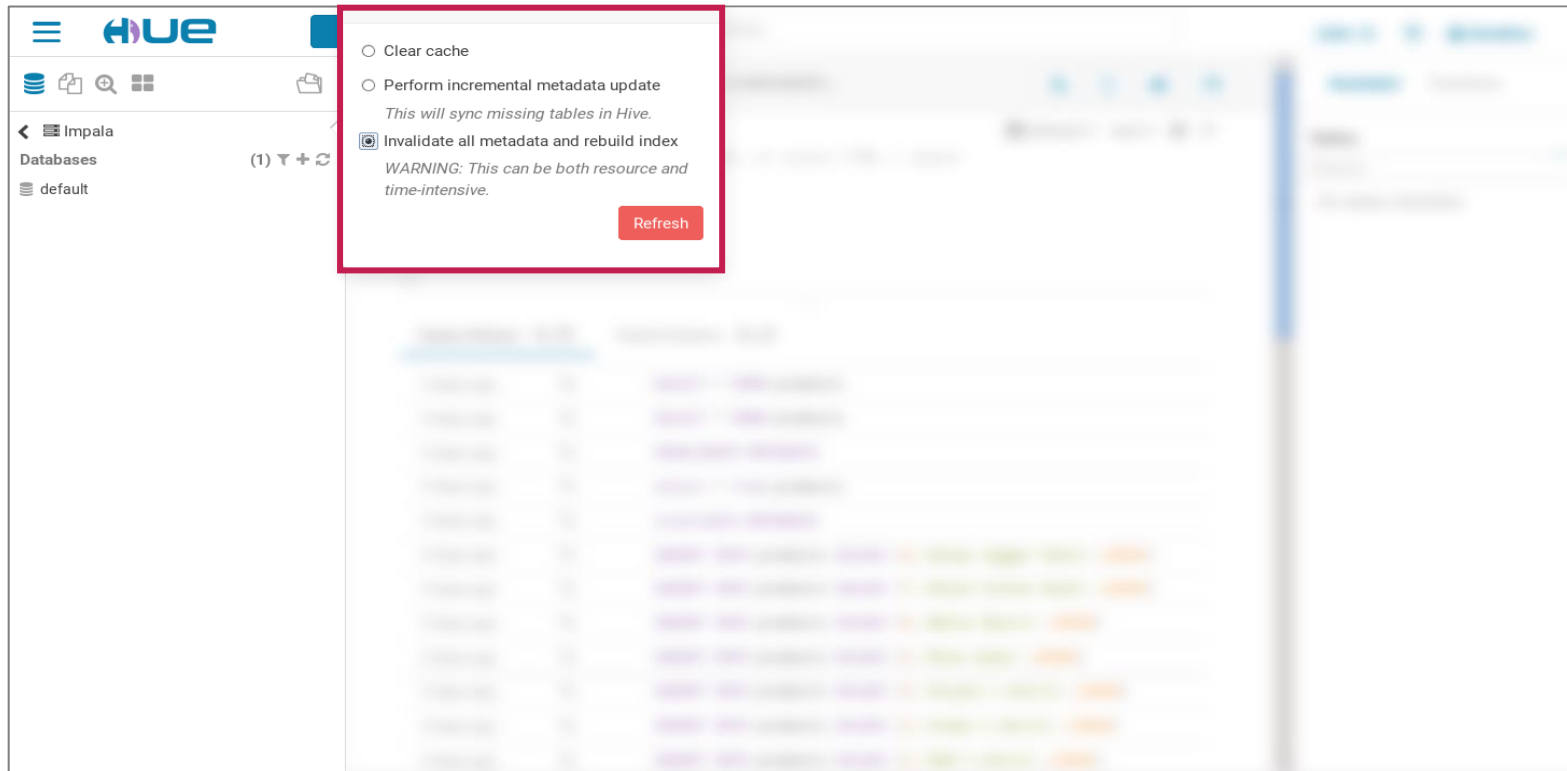
The screenshot displays the Apache Hue interface for the 'airbnb.ab_nyc_2019' table. The left sidebar shows the database 'airbnb' and the table 'ab_nyc_2019' with its schema. The main panel shows the 'Details' tab for the table, displaying a list of columns and their data types, and a 'Sample' tab showing a preview of the table data.

	ab_nyc_2019.id	ab_nyc_2019.name
1	2539	Clean & quiet apt home by the park
2	2595	Skylit Midtown Castle
3	3647	THE VILLAGE OF HARLEM....NEW YORK !
4	3831	Cozy Entire Floor of Brownstone
5	5022	Entire Apt: Spacious Studio/Loft by central parl
6	5099	Large Cozy 1 BR Apartment In Midtown East
7	5121	BlissArtsSpace!
8	5178	Large Furnished Room Near B'way

- Provides detailed information about the clicked table

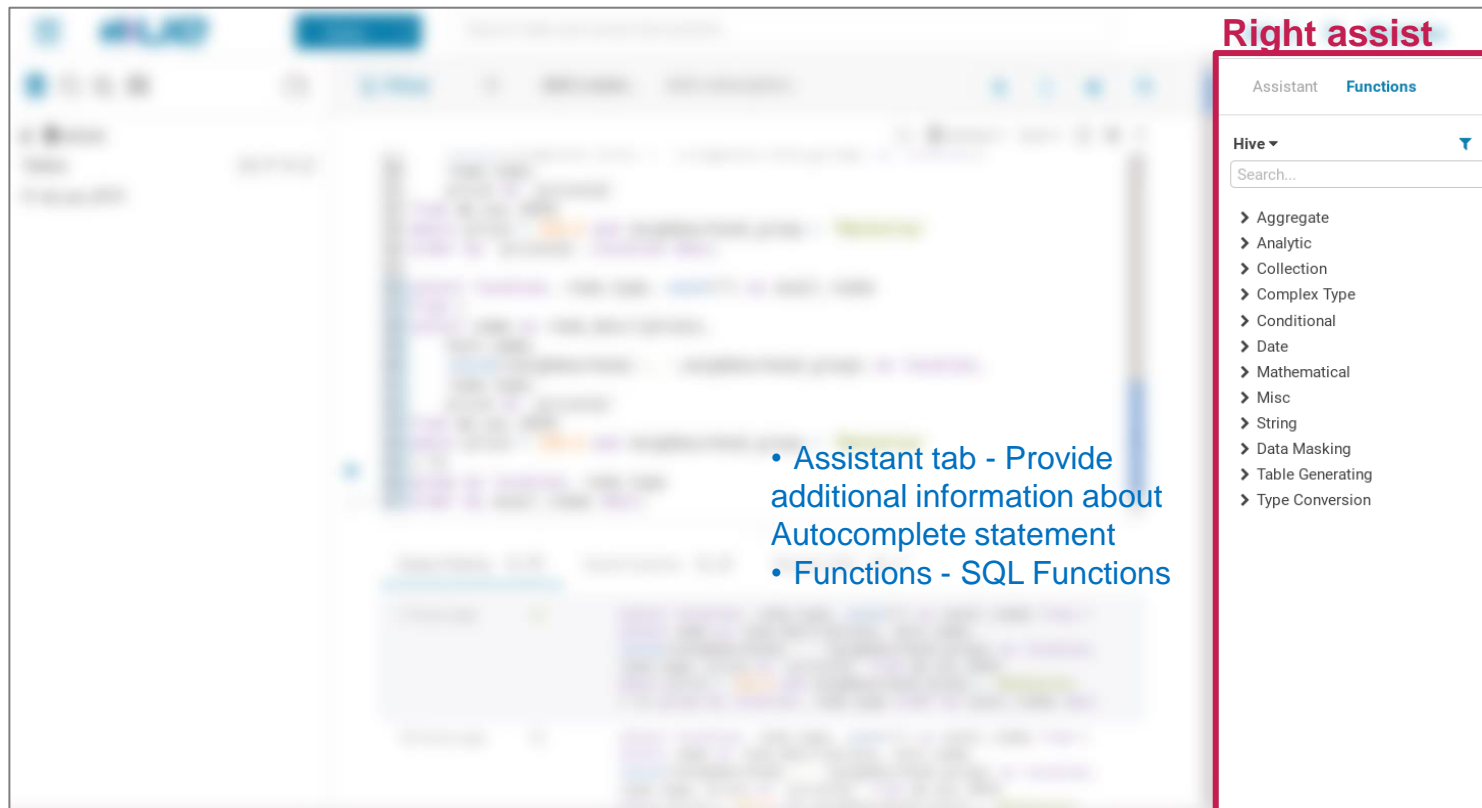
Hue : Truy vấn (4/6)

I Impala - Vô hiệu hóa siêu dữ liệu



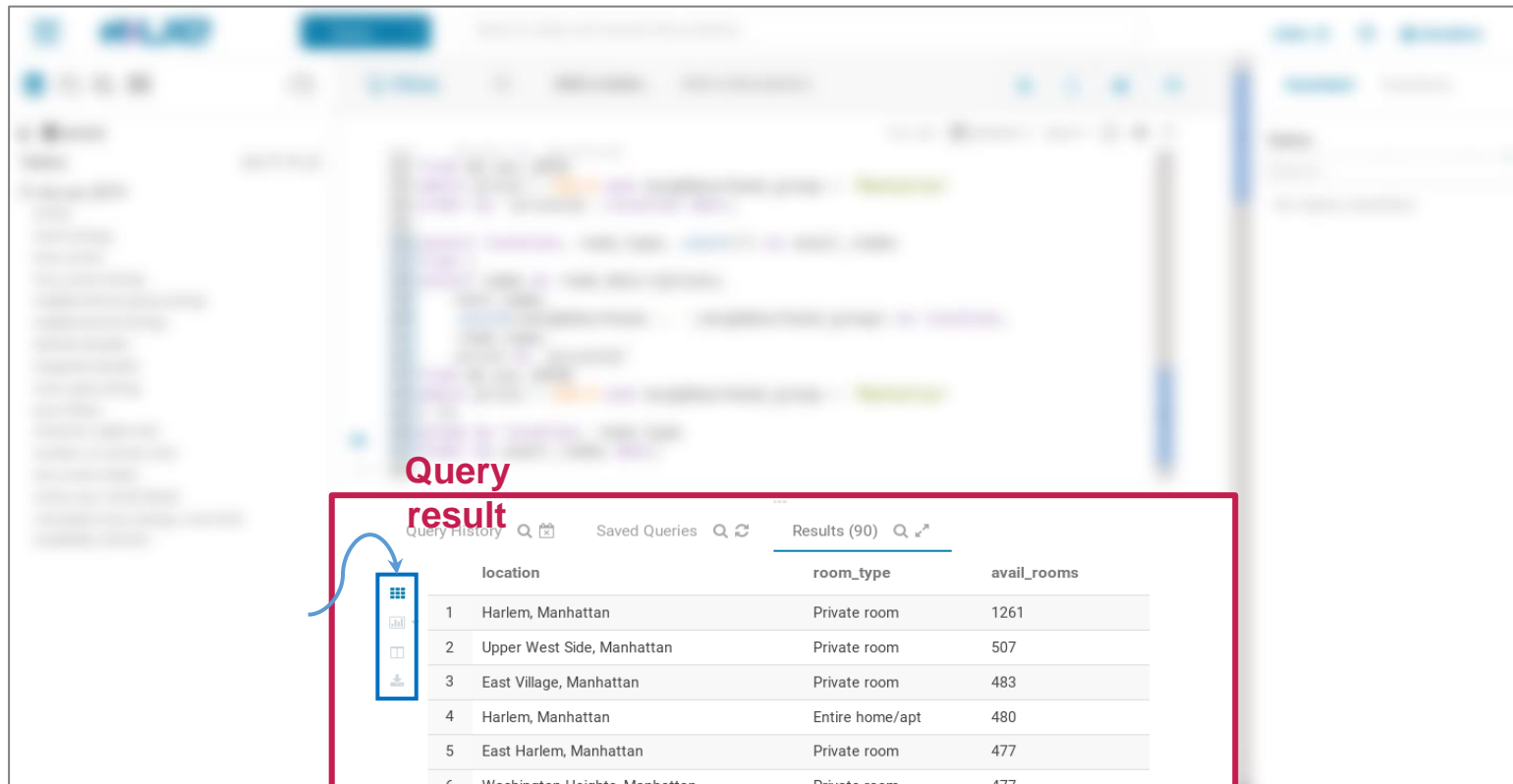
Hue : Truy vấn (5/6)

I Hỗ trợ phải



Hue : Truy vấn (6/6)

I Kết quả truy vấn



Query result

Query History Saved Queries Results (90)

	location	room_type	avail_rooms
1	Harlem, Manhattan	Private room	1261
2	Upper West Side, Manhattan	Private room	507
3	East Village, Manhattan	Private room	483
4	Harlem, Manhattan	Entire home/apt	480
5	East Harlem, Manhattan	Private room	477
6	Washington Heights, Manhattan	Private room	477

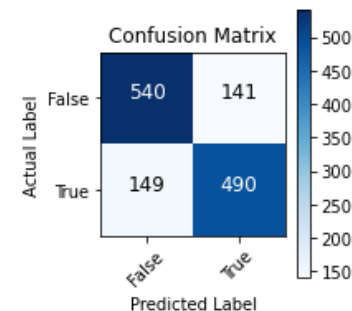
Giới thiệu Jupyter Notebook

- Jupyter Notebook được phát triển từ Ipython Notebook
- Jupyter Notebook là sổ ghi chép điện toán tương tác mã nguồn mở dành cho một số ngôn ngữ lập trình khác nhau bao gồm cả Python
 - Trên thực tế, Jupyter Notebook cũng hỗ trợ Julia, R, Haskell, Ruby, v.v.
- Nó là một ứng dụng dựa trên web; chạy trên trình duyệt
- Một lợi thế lớn của Jupyter là khả năng kết hợp văn bản và mã được định dạng

XGBoost Prediction Result

The confusion matrix shows number of predictions for each of the Correct (Positive and Negative) prediction and Incorrect (Positive and Negative) predictions

```
In [46]: print_confusion_matrix(xgb, x_test, y_test, "xgboost")
```



	precision	recall	f1-score	support
0.0	0.78	0.79	0.79	681
1.0	0.78	0.77	0.77	639
accuracy			0.78	1320
macro avg	0.78	0.78	0.78	1320
weighted avg	0.78	0.78	0.78	1320

Các tính năng của Jupyter Notebook

- I Một số tính năng đáng chú ý của Jupyter Notebook là:
 - ▶ Chức năng tự động hoàn thành
 - ▶ Mã được tổ chức trong các đơn vị tế bào
 - ▶ Hỗ trợ Markdown, LaTeX, v.v. để định dạng văn bản
- I Để chạy sổ ghi chép, hãy chạy lệnh sau tại thiết bị đầu cuối

```
$ jupyter notebook
```

Giới thiệu Jupyter Notebook (1/5)

- Khi bạn bấm vào một ô, đường viền của nó sẽ chuyển sang màu xanh lá cây. Đây là chế độ edit




In []: x=123

Tổ hợp phím	Hành động
CTRL + a	Chọn toàn bộ dòng.
CTRL + d	Xóa toàn bộ dòng.
CTRL + z	Hoàn tác thay đổi cuối cùng.
CTRL + s	Lưu và điểm kiểm tra.

- Trong cả chế độ chỉnh sửa và lệnh, nhấn [SHIFT] + [ENTER] để chạy hoặc biên soạn ô đã chọn

Giới thiệu Jupyter Notebook (2/5)

- Bạn có thể vào chế độ Command bằng cách chọn một ô rồi nhấn phím [ESC]



```
In [ ]: x=123
```

- Một số phím tắt hữu ích khi ở chế độ Command là:

Tổ hợp phím	Hành động
a	Chèn một ô ở trên.
b	Chèn một ô ở dưới.
d + d (twice)	Xóa ô hiện tại.
m	Thay đổi loại ô thành Markdown .
y	Thay đổi loại ô thành Code.

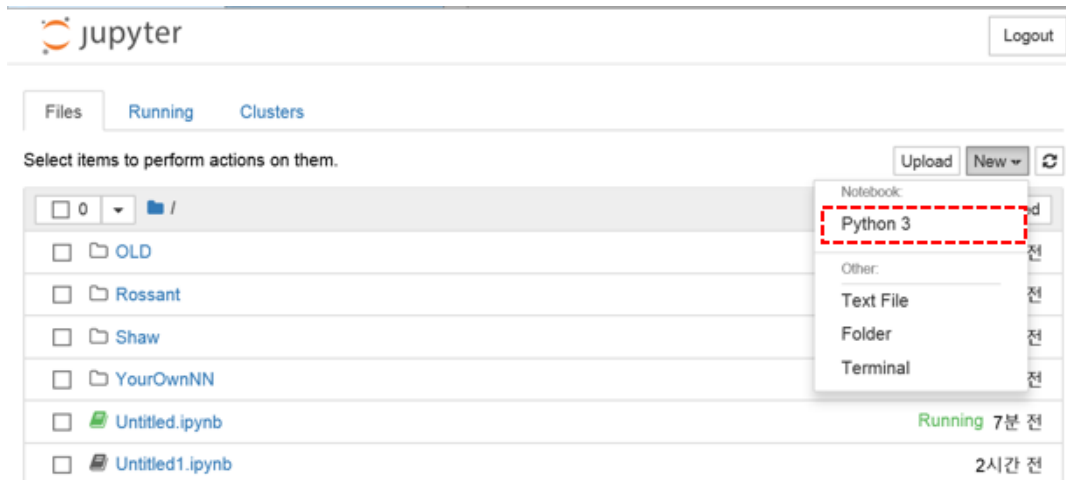
Giới thiệu Jupyter Notebook (3/5)

I Một số thẻ Markdown hữu ích:

Tag	Hành động
#	Tiêu đề lớn nhất.
##	Tiêu đề lớn thứ hai.
-, *, or number	Danh sách vật phẩm
>	Trích dẫn.
[text](URL)	Liên kết.
* <i>Italic</i> *	Phông chữ in nghiêng
** Bold **	Phông chữ in đậm
\$ ~~ \$, \$\$ ~~ \$\$	Biểu thức LaTeX.
-----	Đường chân trời.
 	Ngắt dòng.

Giới thiệu Jupyter Notebook (4/5)

- I Từ Home, nhấp vào tab **Running** để xem tất cả các sổ ghi chép đang chạy
 - ▶ Bạn có thể nhấp vào nút **Shutdown button** để dừng ghi chép
- I Nhấn danh mục **New** rồi chọn **Python 3** để bắt đầu một ghi chép mới



Giới thiệu Jupyter Notebook (5/5)

■ Nhập vào số tay rồi chọn **Cell** → **Run All** để thực hiện liên tiếp tất cả các ô từ trên xuống dưới

Cell	Kernel	Widgets
Run Cells		
Run Cells and Select Below		
Run Cells and Insert Below		
Run All		
Run All Above		
Run All Below		
Cell Type		▶
Current Outputs		▶
All Output		▶

In [27]: # Creating a Logistic Regression Model by fitting the data

```
lr = LogisticRegression(lr=0.01, num_iter=1000, verbose=True)
lr_log = lr.fit(x_train_lr, y_train_lr)
```

```
epoch: 0      loss: 0.6931471803599453
epoch: 10     loss: 0.6904899637675816
epoch: 20     loss: 0.6889239094938908
epoch: 30     loss: 0.6877765367187595
epoch: 40     loss: 0.6868688414532079
epoch: 50     loss: 0.6861264028512316
epoch: 60     loss: 0.685506647126465
epoch: 70     loss: 0.6849811804514382
epoch: 80     loss: 0.6845297410991716
epoch: 90     loss: 0.6841373605861225
epoch: 100    loss: 0.6837927364357145
epoch: 110    loss: 0.6834871959245932
epoch: 120    loss: 0.683213997239752
epoch: 130    loss: 0.682967841058459
epoch: 140    loss: 0.6827445205184289
epoch: 150    loss: 0.6825406655388402
epoch: 160    loss: 0.6823535532062365
epoch: 170    loss: 0.6821809653950383
epoch: 180    loss: 0.6820210807339536
epoch: 190    loss: 0.6818723918967042
```

[Lab1]

Làm việc với Trình soạn thảo truy vấn tại Hue



[Lab2]

Sử dụng chức năng bản đồ trong HUE



Bài 2

Công cụ trực quan hóa dữ liệu thương mại phổ biến

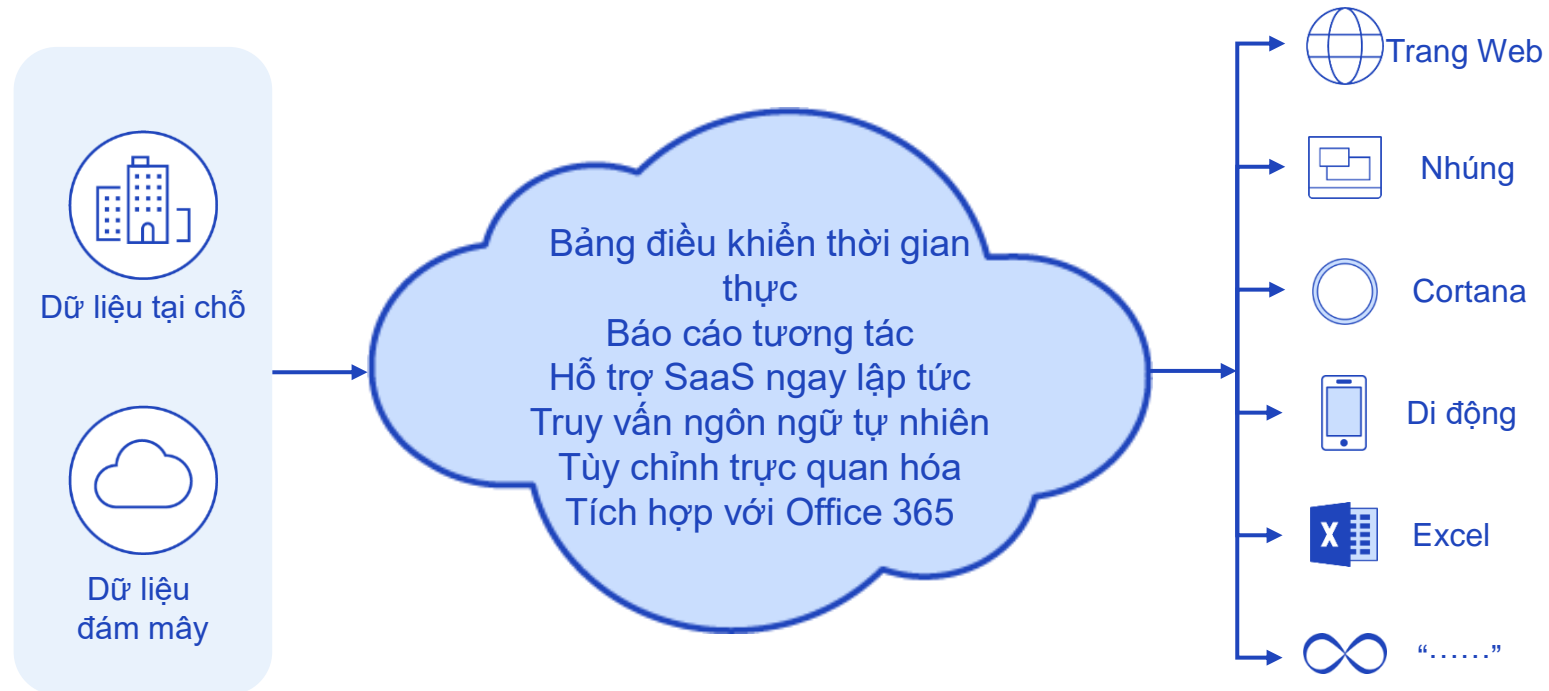
Trực quan hoá dữ liệu

Bài 2

Công cụ trực quan hóa dữ liệu thương mại phổ biến

| 2.1. Giới thiệu về Power BI

Power BI là gì



Công cụ hỗ trợ của Power BI

I Chúng ta có thể làm gì với Power BI?

Tác giả



Power BI
Desktop

Chia sẻ và phối hợp



Dịch vụ
Power BI

Large scale deployments



Power BI
Premium

Share and collaborate



Báo cáo máy chủ
Power BI

App dev



Power BI
Embedded

Phân tích dữ liệu miễn phí
Và công cụ soạn thảo báo cáo

Hiện đại dựa trên đám mây
Giải pháp phân tích kinh doanh

Công suất dành riêng cho
Tăng hiệu suất

Máy chủ báo cáo tại chỗ

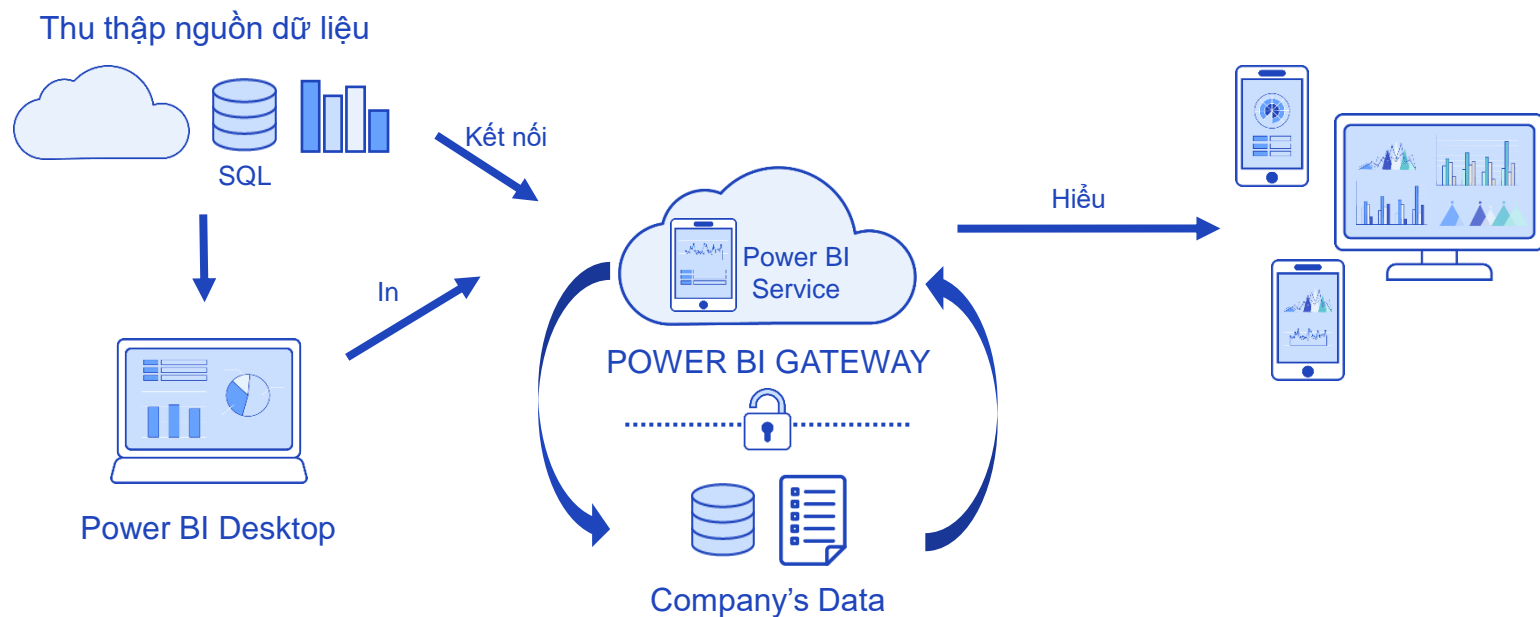
Phân tích trực quan được nhúng
Trong các ứng dụng của bạn

Các thành phần của Power BI

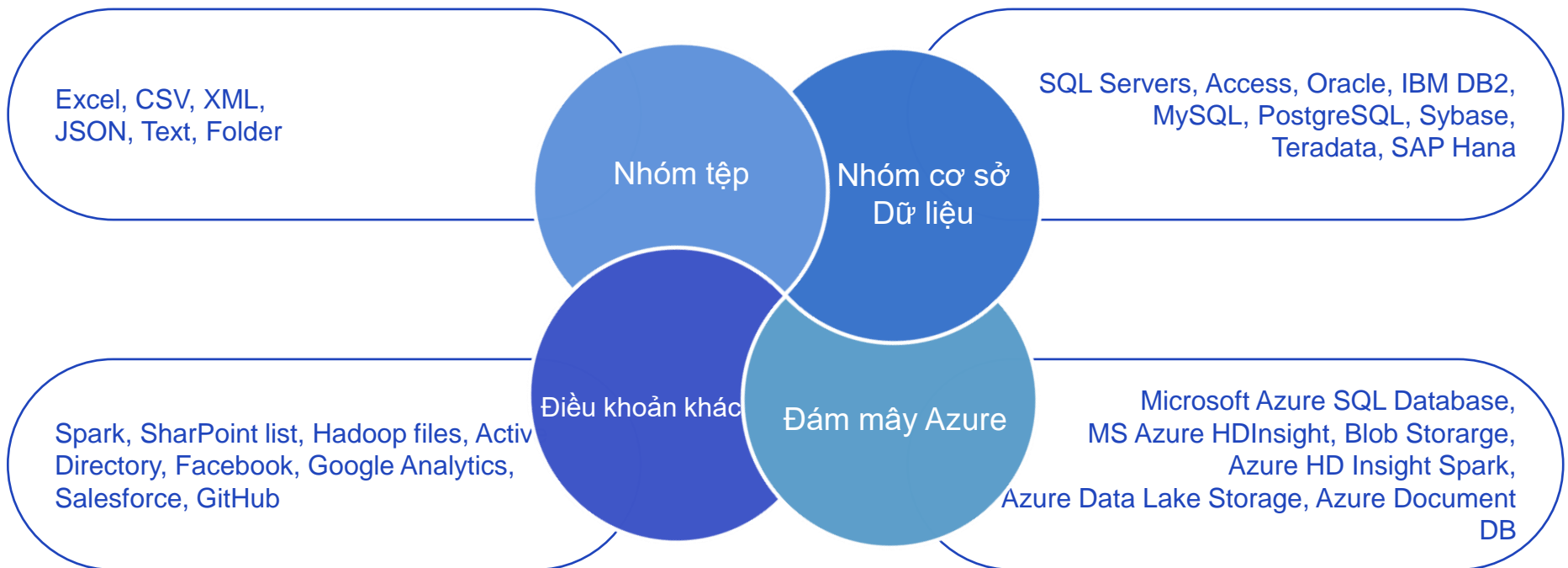
Power Query	Dịch vụ truy cập, tìm kiếm và chuyển đổi dữ liệu từ nhiều nguồn dữ liệu khác nhau. Hỗ trợ kết hợp dữ liệu
Power Pivot	Cung cấp các công cụ cho dữ liệu trong bộ nhớ để mô hình hóa
Power View	Các công cụ để biểu thị dữ liệu bằng hình ảnh và sử dụng chúng để phân tích
Power Map	Các công cụ và khả năng để trực quan hóa dữ liệu không gian địa lý trên các mô hình 3D trong bản đồ
Power Q&A	Tìm kiếm hoặc khám phá thông tin chi tiết từ dữ liệu của bạn bằng cách nhập ngôn ngữ truy vấn tự nhiên

Luồng công việc của dữ liệu và quy trình

I Kiến trúc cơ bản và quy trình công việc của Power BI



Nguồn dữ liệu được Power BI hỗ trợ



Hỗ trợ các công thức DAX và Tập lệnh R

- I DAX là Biểu thức phân tích dữ liệu
 - ▶ Một ngôn ngữ chức năng
- I Một ngôn ngữ công thức được sử dụng trong các dịch vụ phân tích
 - ▶ Bao gồm các hàm, toán tử và giá trị để thực hiện các phép tính và truy vấn nâng cao trên dữ liệu trong các bảng và cột có liên quan trong mô hình dữ liệu dạng bảng

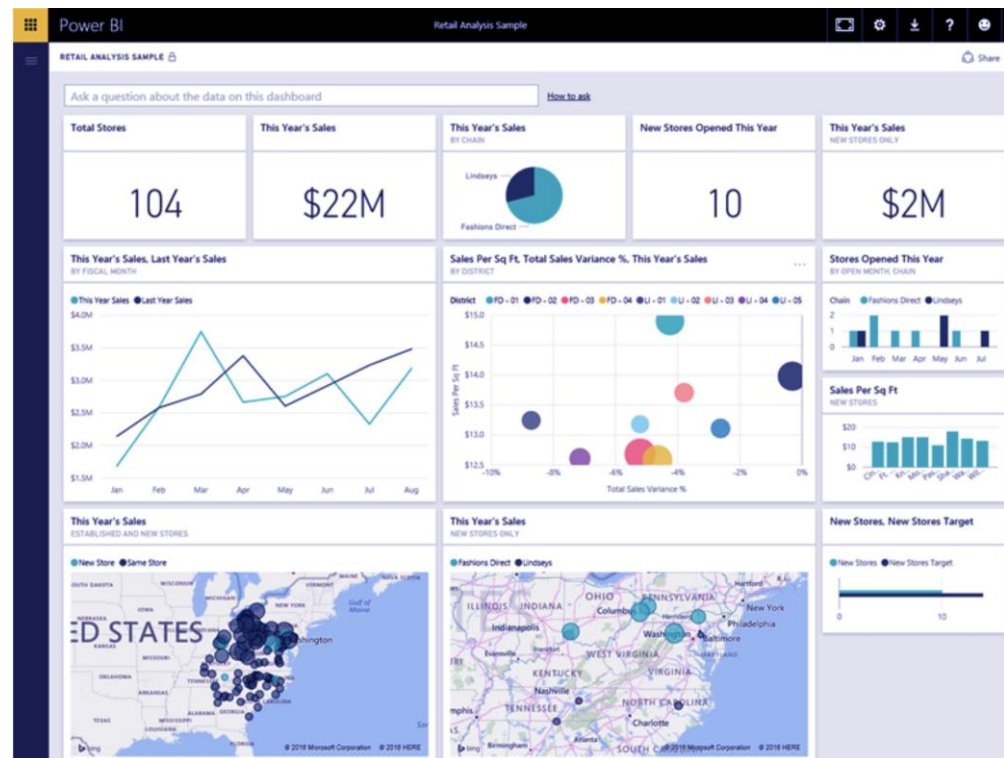
```
ProjectedSales2021 = SUM(Sales[TotalSales2020])*1.21
```

- I R Script có thể được sử dụng để thực hiện chuyển đổi và định hình dữ liệu

```
library(mice)
tempData <- mice(dataset, m=1, maxit=50, meth='pmm', seed=100)
completedData <- complete(tempData, 1)
output <- dataset
output$completedValues <- completedData$"SMI missing values"
```


Khối Căn bản - Trực quan hóa

- Power BI cung cấp nhiều sơ đồ, biểu đồ, bóng đổ và tùy chỉnh
- Các hình ảnh trực quan khác nhau có nhiều tùy chọn tùy chỉnh có thể được sử dụng để làm nổi bật thêm thông tin cốt lõi cần được truyền tải



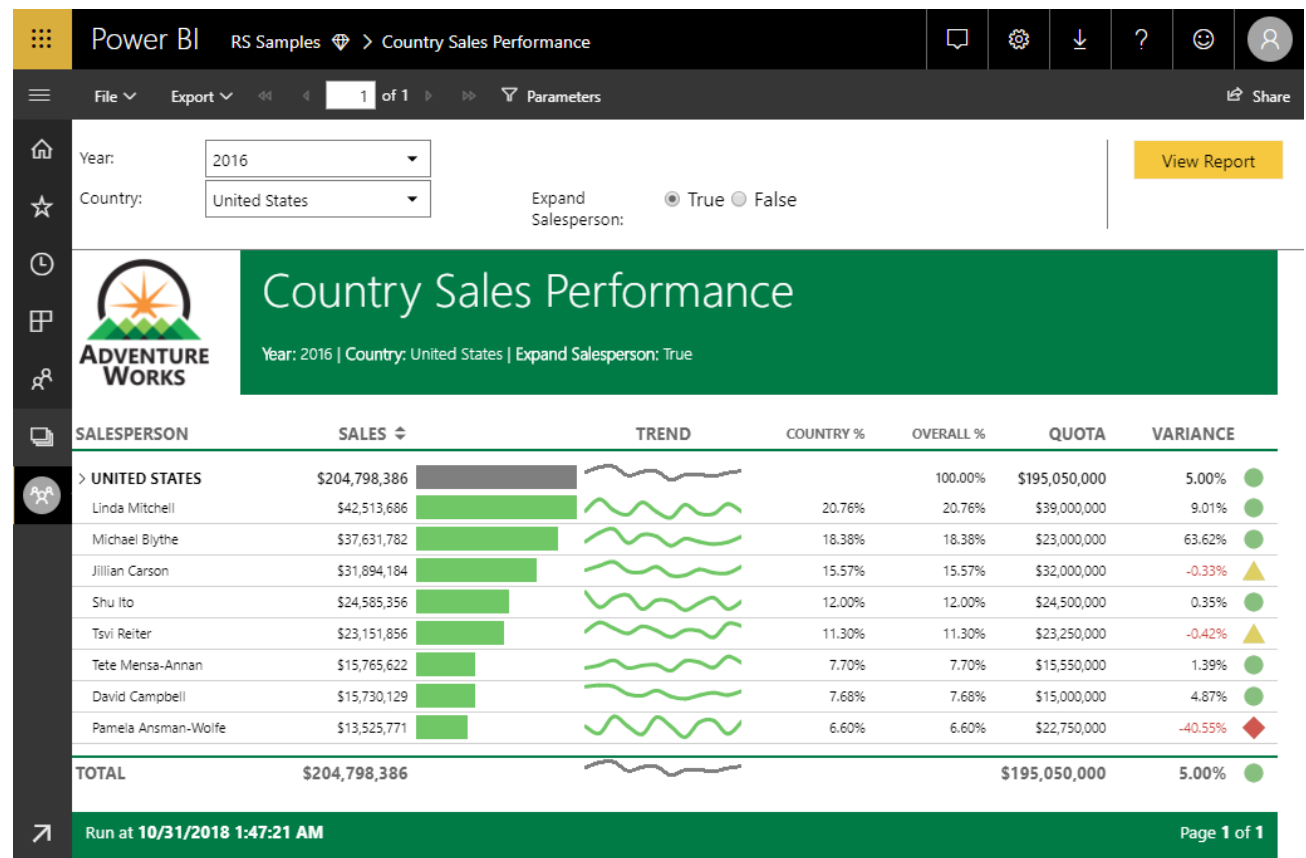
Khối Căn bản - Bộ dữ liệu

- Bộ dữ liệu là dữ liệu từ các nguồn dữ liệu được hỗ trợ khác nhau đã được nhập, làm sạch và sẵn sàng để phân tích.
- Chúng ta đã thấy nhiều loại nguồn dữ liệu mà từ đó Power BI có thể tạo Bộ dữ liệu
- Hiển thị ở đây là một bộ dữ liệu dựa trên Excel

	B	C	D	E	F	G	H
1	Year	Month	Month Name	Calendar Month	Births	Births Per Day	Births (Normalized)
2119	2004	1	January	1/1/2004	2,937	94.7	2842
2120	2004	2	February	2/1/2004	2,824	97.4	2921
2121	2004	3	March	3/1/2004	3,128	100.9	3027
2122	2004	4	April	4/1/2004	2,896	96.5	2896
2123	2004	5	May	5/1/2004	3,008	97.0	2911
2124	2004	6	June	6/1/2004	3,047	101.6	3047
2125	2004	7	July	7/1/2004	2,981	96.2	2885
2126	2004	8	August	8/1/2004	3,079	99.3	2980
2127	2004	9	September	9/1/2004	3,219	107.3	3219
2128	2004	10	October	10/1/2004	3,547	114.4	3433
2129	2004	11	November	11/1/2004	3,365	112.2	3365
2130	2004	12	December	12/1/2004	3,143	101.4	3042
2131	2005	1	January	1/1/2005	2,921	94.2	2827
2132	2005	2	February	2/1/2005	2,699	96.4	2892
2133	2005	3	March	3/1/2005	3,024	97.5	2926
2134	2005	4	April	4/1/2005	3,037	101.2	3037
2135	2005	5	May	5/1/2005	3,231	104.2	3127
2136	2005	6	June	6/1/2005	3,163	105.4	3163
2137	2005	7	July	7/1/2005	3,119	100.6	3018
2138	2005	8	August	8/1/2005	3,156	101.8	3054
2139	2005	9	September	9/1/2005	3,439	114.6	3439

Khối Căn bản - Báo cáo

- Các báo cáo cho phép tổ chức nhiều hình ảnh trực quan lại với nhau thành các báo cáo hợp lý
 - Các báo cáo có thể mở rộng trên nhiều trang mà sau đó có thể được truy cập bằng các tab
- Báo cáo có thể tương tác nơi chế độ xem có thể đặt tham số cho báo cáo



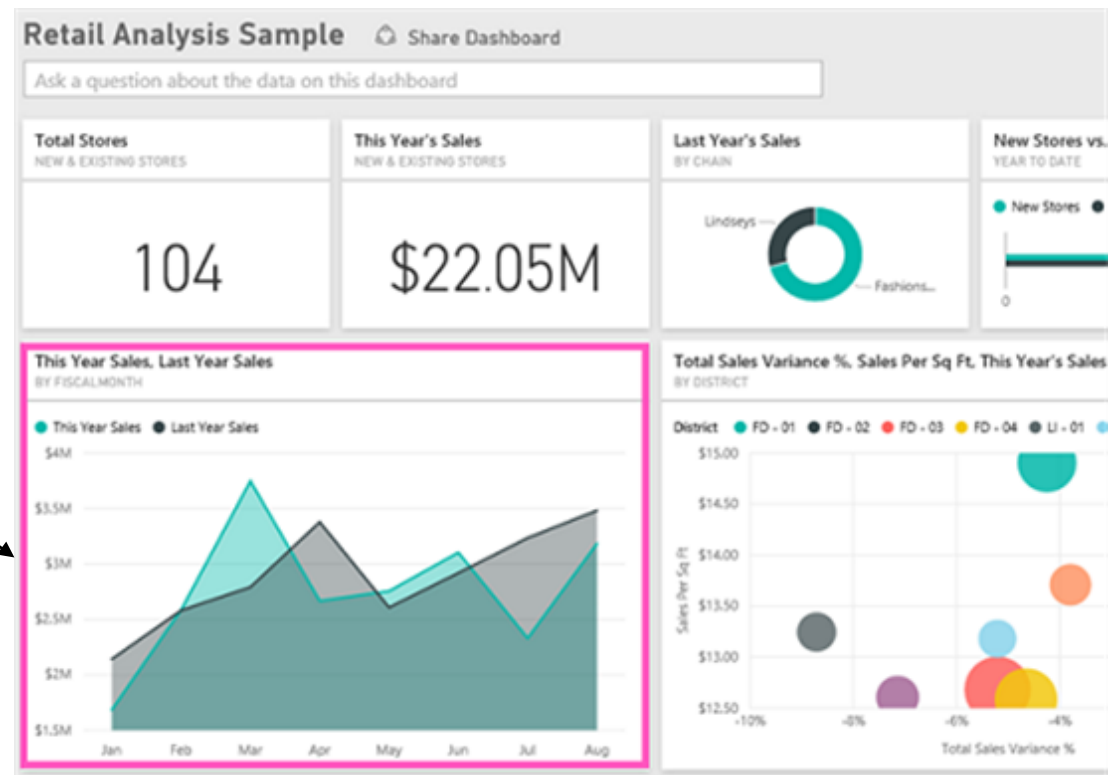
Khối Căn bản - Bảng điều khiển

- Trang tổng quan có thể được coi là các trang cấp cao mà từ đó khách hàng có thể đi sâu hơn vào các báo cáo khác nhau
- Nói chung, Báo cáo cung cấp thông tin chi tiết hơn trong khi Bảng điều khiển được sử dụng để cung cấp các thông tin chính



Khối Căn bản - Đường

- Đường là các biểu đồ và đồ thị cung cấp một mẫu thông tin duy nhất
- Báo cáo và Bảng điều khiển thường bao gồm tập hợp nhiều Đường để thu thập theo cách hợp lý để tạo Báo cáo chi tiết cuối cùng về một chủ đề cụ thể



Power BI và so sánh Tableau

I Sự khác nhau giữa Power BI và Tableau

Power BI	Tableau
Power BI hoạt động tốt nhất với dữ liệu có kích thước hạn chế. Làm việc với dữ liệu có kích thước lớn hơn sẽ yêu cầu các bước sắp xếp và tóm tắt dữ liệu thành các kích thước dễ quản lý hơn	Tableau có thể xử lý khối lượng dữ liệu khổng lồ với hiệu suất tuyệt vời
Power BI rất dễ bắt đầu	Tableau có đường cong học tập dốc hơn
Power BI được sử dụng bởi cả người dùng có kinh nghiệm và người dùng mới, thường để có thông tin chi tiết nhanh chóng và dễ dàng	Tableau được sử dụng bởi các nhà phân tích và người dùng có kinh nghiệm với mục đích trực quan hóa công việc phân tích của họ
Power BI cung cấp nhiều điểm dữ liệu để trực quan hóa	Tableau cung cấp nhiều chức năng trực quan hóa dữ liệu
Power BI có ít trình kết nối hơn với nguồn dữ liệu	Tableau có nhiều trình kết nối với nhiều loại nguồn dữ liệu khác nhau
Power BI là tốt nhất cho các công ty vừa và nhỏ. Mặc dù Power BI Premium rất phù hợp cho tổ chức lớn hơn với nhu cầu lớn hơn.	Tableau phù hợp nhất cho các doanh nghiệp vừa và lớn
Hỗ trợ khách hàng hạn chế	Hỗ trợ khách hàng rộng rãi

[Lab3]

Làm việc với Power BI





SAMSUNG

Together for Tomorrow!
Enabling People

Education for Future Generations

©2021 SAMSUNG. All rights reserved.

Samsung Electronics Corporate Citizenship Office holds the copyright of book.

This book is a literary property protected by copyright law so reprint and reproduction without permission are prohibited.

To use this book other than the curriculum of Samsung innovation Campus or to use the entire or part of this book, you must receive written consent from copyright holder.