

# AZURE SYNAPSE STUDIO DEVELOPMENT TOOLS

---

RUSS LOSKI



# INTRODUCTION

---

- Russ Loski
- Data Engineer
- Husband, dad and grandad
- [russloski@sqlmovers.com](mailto:russloski@sqlmovers.com)
- [www.sqlmovers.com](http://www.sqlmovers.com)
-  <https://twitter.com/sqlmovers>
-  <https://www.linkedin.com/in/russloski>
- Slides and Code: <https://github.com/rloski-public/Presentations/tree/main/Atlanta%202023>

# AGENDA

---

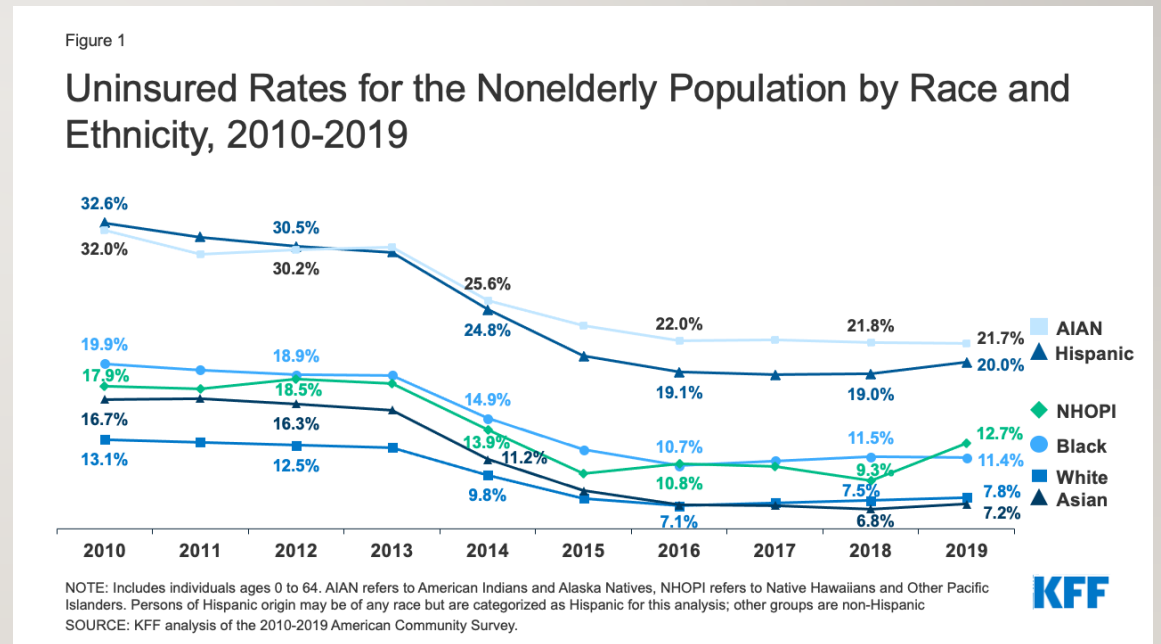
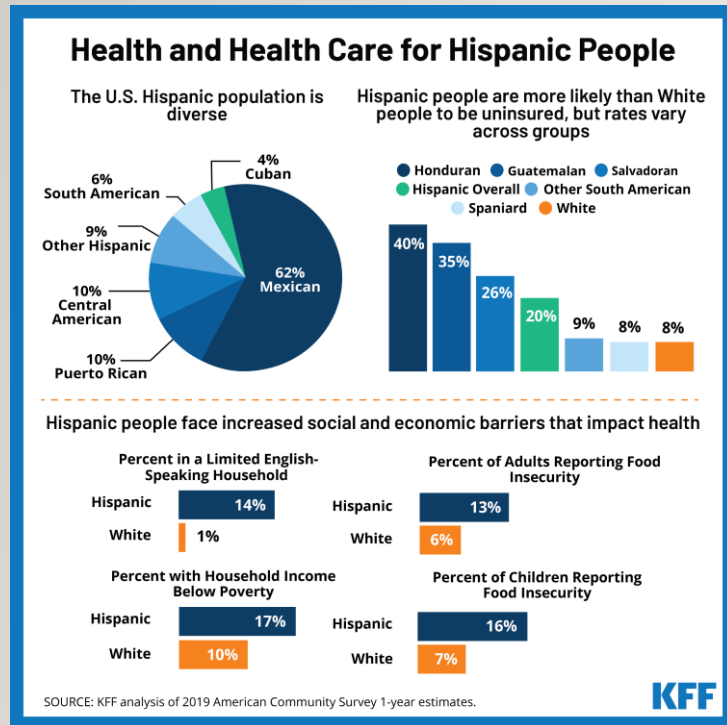
- Data problem
- Overview of the interface
- Developing SQL
- Developing Spark



DATA IS THE NEW TREASURE



# HEALTH INSURANCE RATES FOR HISPANICS



[COW-Hispanic-Hertiage-Month\\_v5.png \(4500×4500\) \(kff.org\)](#)

[Health Coverage by Race and Ethnicity, 2010-2019 | KFF](#)

# PUBLIC DATA SOURCE

---

- American Community Surveys (ACS)
  - [American Community Survey \(ACS\) \(census.gov\)](https://www.census.gov/programs-surveys/acs/data.html)
- Public Use Microdata Sample (PUMS)
  - [American Community Survey Microdata \(census.gov\)](https://www.census.gov/programs-surveys/acs/data/pums.html)
- PUMS 2017
  - <https://www2.census.gov/programs-surveys/acs/data/pums/2017/1-Year/>
    - csv\_hus.zip (Housing file)
    - csv\_pus.zip (Population or Person file)

# WHAT DO I NEED FROM THIS FILE?

---

- Is Hispanic?
- Is insured?
- Population
- Age

# DATA WAREHOUSE

---

- Requires IT involvement
- Requires setting up server and database
- Requires ETL pipeline
- Maybe data not suitable
- Maybe question is only relevant a short time



# OTHER OPTIONS

---

- Power BI (Power Query)
- Excel
- Python
- Databricks
- Azure Synapse Analytics

Microsoft Azure | Synapse Analytics | loskisyapse

Synapse live | Validate all | Publish all

Home | Data | Develop | Integrate | Monitor | Manage

Develop

Filter resources by name

SQL scripts 10

- BuildCensusScriptHealthInsurance
- BuildCensusScriptRooms
- CensusPumsHousing2012
- Create Lookup tables
- SQL script 1
- SQL script 2
- SQL script 3
- SQL script 4
- SQL script 5
- Test Script

Notebooks 1

BuildCensusScriptH...

Run | Undo | Publish | Query plan | Connect to Built-in | Use database test1

```
1 USE test1;
2 SELECT
3     TOP 100 *
4 FROM
5     OPENROWSET(
6         BULK 'https://loskisyapsedatalake.dfs.core.windows.net/loskisyapsefilesystem/census/PUMS/1-year/2017/Population/psam_pusa.csv',
7         FORMAT = 'CSV',
8         PARSER_VERSION = '2.0'
9     ) AS [result];
10
11
12
13
14 SELECT
15     TOP 100 *
16 FROM
17     OPENROWSET(
18         BULK 'https://loskisyapsedatalake.dfs.core.windows.net/loskisyapsefilesystem/census/PUMS/1-year/2017/Population/psam_pusa.csv',
19         FORMAT = 'CSV',
20         PARSER_VERSION = '2.0',
21         HEADER_ROW = TRUE
22     ) AS [result];
23
24 -- Change the data types and focus on the columns of interest
```

# OVERVIEW OF THE INTERFACE

Develop

Filter resources by name

SQL scripts10

BuildCensusScriptHealthInsurance

BuildCensusScriptRooms

CensusPumsHousing2012

Create Lookup tables

SQL script 1

SQL script 2

SQL script 3

SQL script 4

SQL script 5

Test Script

Notebooks2

Demonstration Notebook

Demonstration Notebook Population

BuildCensusScriptH... x

Demonstration Note...

Run

Undo

Publish

Query plan

Connect to

Built-in

Use database

test1

12

-- Make the first row the header

13

14

SELECT

15

...TOP 100.\*

16

FROM

17

...OPENROWSET(

18

...BULK 'https://...dfs.core.windows.net/.../census/PUMS/1-year/2017/Population/psa...

19

...FORMAT='CSV',

20

...PARSER\_VERSION='2.0',

21

...HEADER\_ROW=TRUE

22

...AS [result];

23

24

-- Change the data types and focus on the columns of interest

25

\*\*

Results

Messages

View

Table

Chart

Export results

Search

RT	SERIALNO	DIVISION	SPORDER	PUMA	REGION	ST	ADJINC	P
P	2017000000016	6	1	2500	3	1	1011189	2
P	2017000000031	6	1	1800	3	1	1011189	4
P	2017000000061	6	1	2400	3	1	1011189	1

00:00:11 Query executed successfully.

DEVELOPING SQL



# PROBLEM

---

What is in the PUMS population file? Can it answer my question about Hispanic health insurance rates?



# SERVERLESS SQL POOLS

---

- Can connect to data in data lake
- Can write T-SQL queries
- Pay for the compute when you run query
- Still pay for storing data in the data lake

# ADDITIONAL BENEFITS

---

- Can connect to open datasets
  - [Tutorial: Analyze Azure Open Datasets in Synapse Studio - Azure Synapse Analytics | Microsoft Learn](#)
- Can be used as data source in Power BI
  - [Tutorial: Connect serverless SQL pool to Power BI Desktop & create report - Azure Synapse Analytics | Microsoft Learn](#)

Synapse live

Validate all

Publish all

Develop

Filter resources by name

SQL scripts 10

Notebooks 2

Demonstration Notebook

Demonstration Notebook Population

BuildCensusScriptHe...

Demonstration Note...

Demonstration Not...

Cancel all

Undo

Publish

Outline

Attach to Spark1

Language PySpark (Python)

Variables

Please wait a few minutes while your session starts.

### Demonstration

This is code to demonstrate how to work with the **Azure Synapse Studio** notebooks. You can get more help at [Create, develop, and maintain Synapse notebooks in Azure Synapse Analytics](#).

We will examine the following:

1. Markdown cells [Markdown for Jupyter notebooks cheatsheet](#)
2. Working with the file magic commands
3. Using the mssparkutil class
4. Access file as dataset
5. Display options
  1. Show column information
  2. Show sample data
  3. Display in charts
6. Using multiple languages
7. Using SQL

We won't go deep into how to actually develop. My goal is to get you up and started with some of the mechanics of working with the notebooks.

1 print("Hello world!")

\* 1 min 21 sec - Starting Apache Spark session

+ Code

+ Markdown

### File magic commands

The first thing I like to do is to get my bearings in the file system. There are enough places that I am going to mess up. I need to make sure that the file is where I think that it is.

These are some of the commands that **might** be available in Azure Synapse Studio. <https://ipython.readthedocs.io/en/stable/interactive/magics.html>.

Try %lsmagic to get a full list of the commands.

### lsmagic

This provides a list of the magic commands with a link to the ipython document mentioned above.

Properties

General Related (0)

Name \*

Demonstration Notebook Population

Description

This is a notebook for demonstrating how to work with Azure Synapse Analytics notebooks.

Type

.ipynb notebook

Size

703,404 bytes

Notebook settings

☒ Include cell output when saving

☐ Enable unpublished notebook reference

Session

Configure session

# DEVELOPING SPARK



# PROBLEM

---

I am more comfortable using Python to work with data.



# RESOURCES

---

- [How to use Synapse notebooks - Azure Synapse Analytics | Microsoft Docs](#)
- [SQL scripts in Synapse Studio - Azure Synapse Analytics | Microsoft Docs](#)
- [Markdown for Jupyter notebooks cheatsheet - IBM Documentation](#)
- [Index of /programs-surveys/acs/data/pums/2017/1-Year \(census.gov\)](#)

# CONTACT

---

- Russ Loski
- [russloski@sqlmovers.com](mailto:russloski@sqlmovers.com)
- [www.sqlmovers.com](http://www.sqlmovers.com)



- <https://twitter.com/sqlmovers>



- <https://www.linkedin.com/in/russloski>
- Slides and Code: <https://github.com/rloski-public/Presentations/tree/main/Atlanta%2022023>