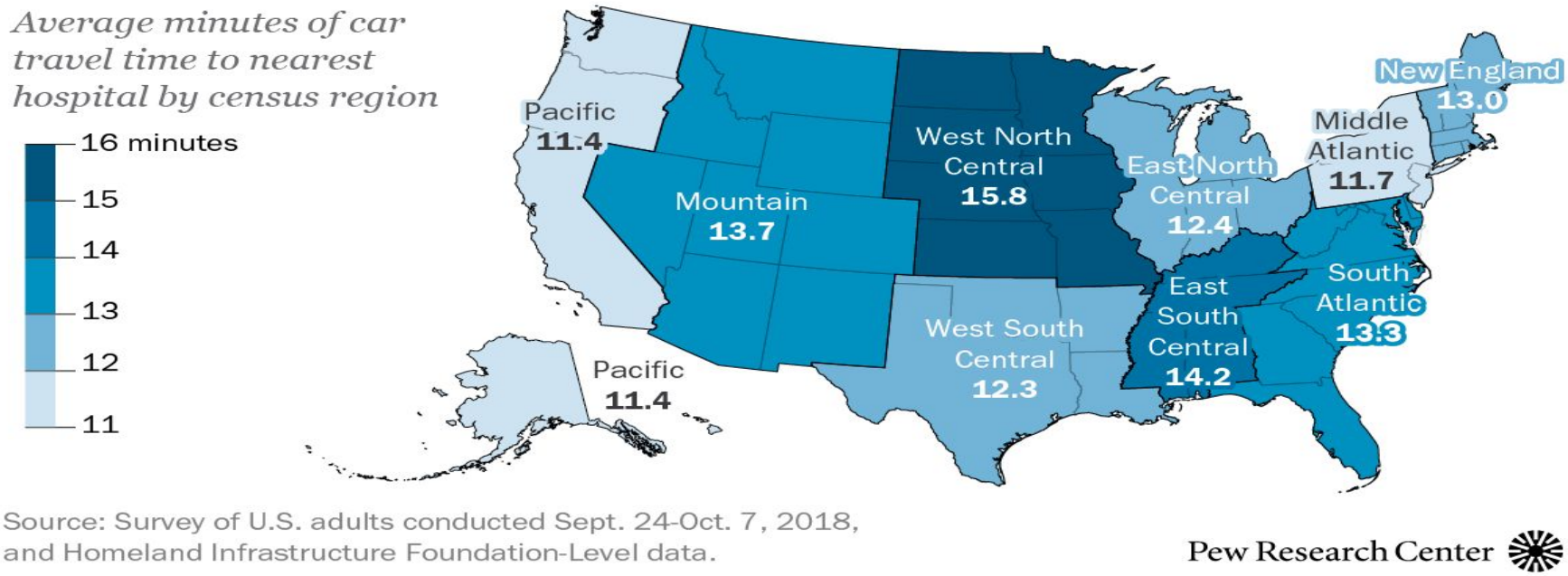
Impact of Nearby Amenities on Airbnb Listings: A study on Big Data Analytics using Spark

Krishna Pothugunta
Eli Broad College of Business
pothugun@broad.msu.edu

Mission

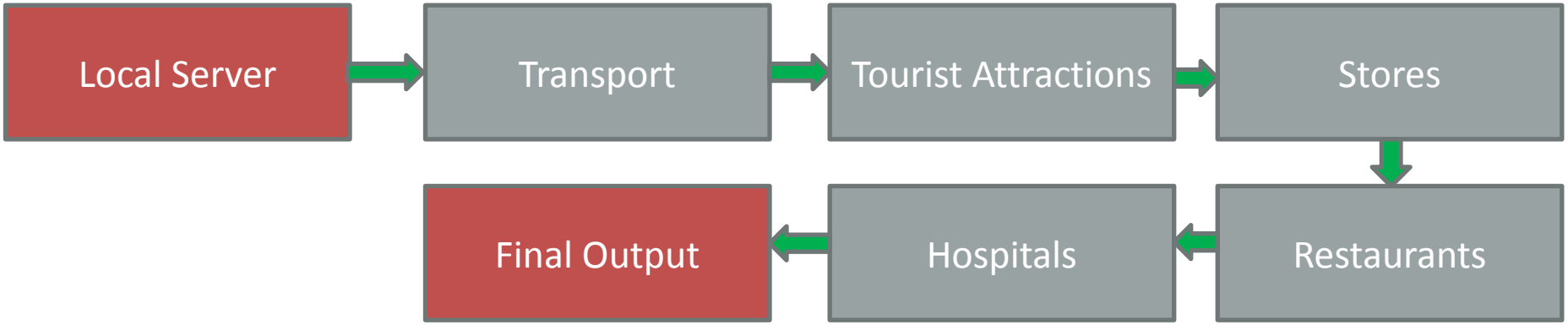
- Calculating the distance between each Airbnb listing and a store for every zip code / borough unit in a designated city
- WHY AIRBNB?
- Post – 2013, the Sharing Economy (SE) platforms have grown tremendously
- Estimated to grow from $14 billion in 2014 to $335 billion by 2025
- Airbnb and Uber are two major players in the SE sector
- What are the consumer preferences in the selection of Airbnb listing?
- Do the nearby amenities play any role in the survival of the Airbnb listing?
  - Transport, Tourist Attractions, Stores, Restaurants and Hospitals

People in Urban region tend to stay closer to hospitals – (Pre- covid)

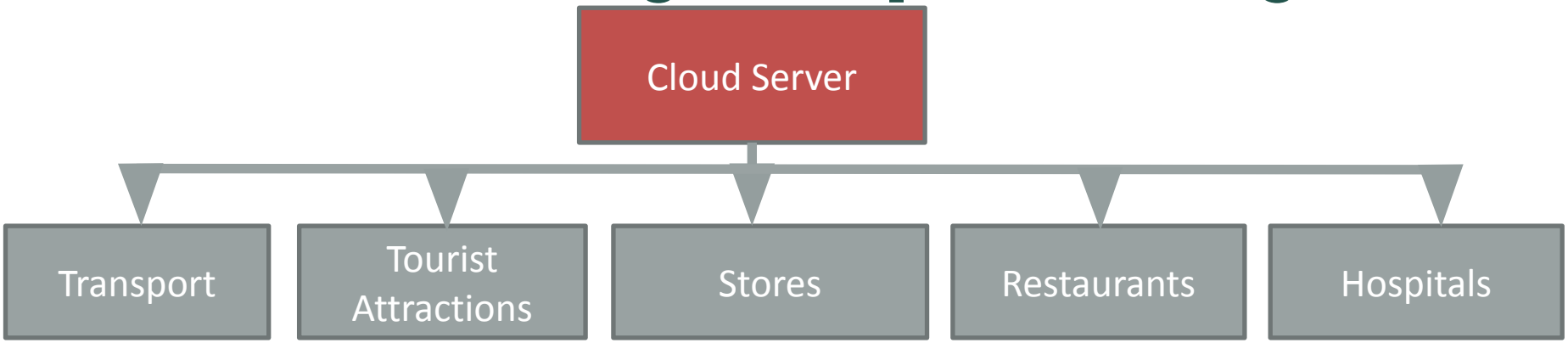


- Will the preference be the same during covid?

Why Cloud Services?



Run the five jobs parallelly



List of Cloud Services Used

1. Storage
   a) Containers
   b) Blob storage
2. Azure Databricks
   a) Clusters
   b) Notebooks

- Experiment design
  1. A single cluster for five separate notebooks
  2. Programming language
     a) Python
     b) SQL

	Airbnb	Hospitals	Total
Bronx	1474	851	1254374
Brooklyn	20086	2954	59334044
Manhattan	25032	7862	196801584
Queens Village	6967	41	285647
Staten Island	429	808	346632

TOTAL = 258M

Importance of CLOUD

- **Cluster Employed: 42 GB, 12 Cores, DBR 10.4 LTS, Spark 3.2.1**
- **Number of nodes running = 3 (STANDARD 14 GB Memory, 4 Cores)**

	PANDAS(HOURS)	SQL(SECONDS)	EFFICIENCY
Transport	0.63	6.33	35729%
Tourist Spots	4.16	17	87994%
Hospitals	30	90	119900%
Stores	42	120	125900%
Restaurants	50	130	138362%

- SIMPLY WOW
- ON A LOCAL MACHINE – TIME SPENT = 50+42+30+4.16+0.63 ~ 5 days
- INTRO VIDEOS ON AZURE SPARK (MEDIASPACE) – GOOD STARTING POINT

Gained Wisdom

- Closing the notebook does not halt the runtime
  - Cluster must be active (No stopping or detaching the cluster manually during the runtime)
- Interaction between spark and python / SQL programming scripts
  - Spark reads any data with column values designated as STRING values (unless schema is specified)
  - Always be careful when the values are converted to Integer / Double / Float (i.e., using cast())
  - LATITUDE
    - INTEGER (40)
    - DOUBLE (40.75356)
    - FLOAT (40.753559)
  - Makes a huge amount of difference in calculation of distance

Key Insights

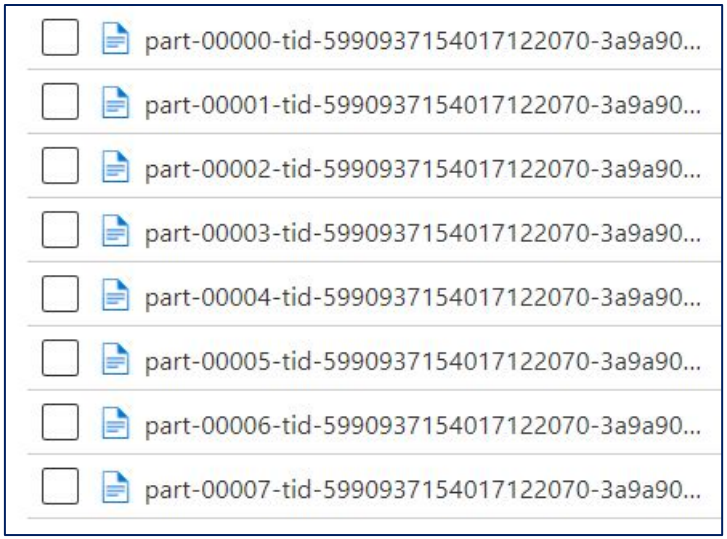
- Not much difference in processing time between having a single / multiple clusters
  - Going for a cluster with higher processing capacity increases the cost drastically
- Data frames can be interchanged between Pandas and Spark
  - Use df.toPandas() and spark.createDataFrame(df)
- Read the data from CSV files using Python
- Replace for loop (iterations) in Pandas with JOINS and VIEWS using SQL

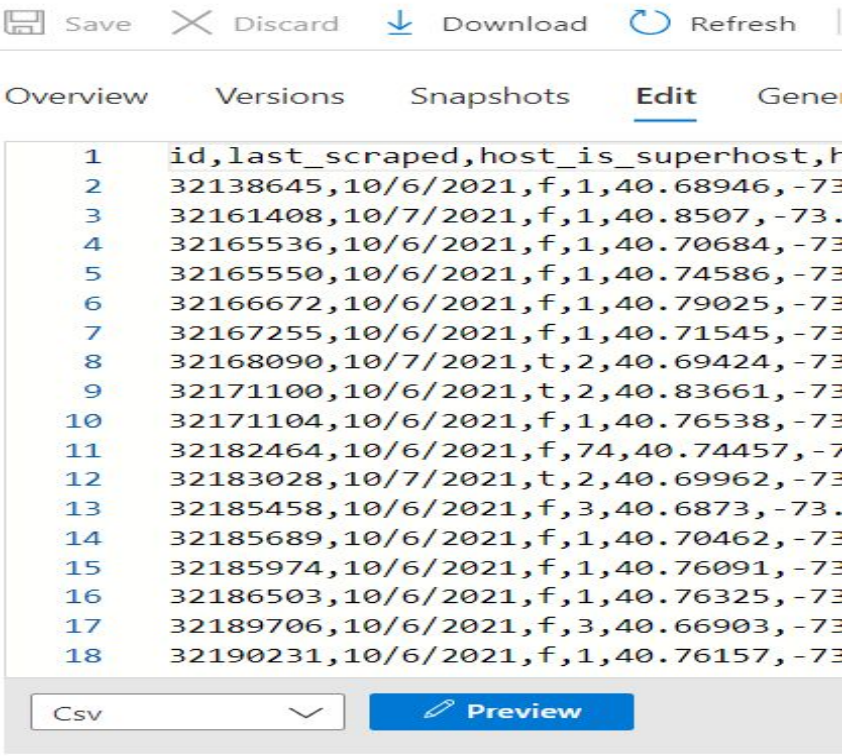
- Databricks Spark is very user friendly
  - Can change into your language of preference (Python, R, Scala and SQL) within the notebook
  - Charts are in – built with display() or show() function

WRITING RESULTS TO CSV FILES

- Must initialize SQL functions to assign Schema
  - Incase of any columns which have NULL values (ERROR)
- Use Final_transport.write.csv()

- Not So Straightforward
- Retrieving the results is tough, when the commands are not clear





MAJOR TAKEAWAYS

- Use the DOWN arrow next to download option
- Can only download up to 1 million rows
- Filter function in the CSV file shows a warning sign

- Learn more on the business applications of SPARK through various programming language interfaces
- In the present project:
  - **a)** Adding more conditions in the for loop created more burden on processing speed
  - **b)** SQL filters the tables easily by WHERE and GROUPBY

WHY NOT JUST USE SQL?

- By 2025, it's estimated that 463 exabytes of data will be created each day globally (Jeff Desjardins 2019)
  - Equivalent to 212 Million DVDs per day

- Present project is strictly limited to NY city
  - File sizes are less than 10 MB

- Airbnb demands physical presence of the consumer to enjoy the service

- What would happen if the research question changes to online tutoring?
  - Distance does not really matter at all
  - Whole world is the market now

Conclusion

- Use cloud services
  - Programming logic is repeated across several records
  - Resources of local machine can be optimized
- Expanding the work to other cities
  - Compare NY survival rates with those of LA, Chicago
  - Rural Areas (Jump in Airbnb listings posted after COVID)
- Create a NEW BLOG about the application of Databricks Spark soon

HIOX

Thank You Any Questions