**FINAL PROJECT**

Analytics Team

Advanced Data Analytics, University of North Texas

ADTA 5240: Harvesting, Storing and Retrieving data

Dr. Tony Fantasia

May 5, 2024

# **THE SIGNIFICANCE OF THE DATA LIFECYCLE: AN ILLUSTRATION USING THE DALLAS PD DATA​​**

**TEAM MEMBERS**

* Hasritha Polamarasetti - 11706930
* Nitesh Runku - 11736022
* Dharmendra Basava - 11654254
* Lalith Golla - 11694964
* Akhila Uppala - 11661864

# **SUMMARY**

Data is here to stay and when leveraged upon by Businesses, governments and individuals, it can transform life to unimaginable dimensions of predictability and amazing possibilities. Security as a social science has increasingly been integrating Data in order to create a proactive approach to security threats. To underscore these assertions, a public safety dataset from the Chicago police department was sourced from the Dallas open data portal. This is the first step of the data life cycle which is data generation. The data acted as a foundation stone to this analysis that sought to illustrate the Data Lifecycle while underlining its importance. The second phase of the data life cycle which is data collection stands here, where the data is collected.

In accordance to the data Lifecycle steps, the data was preprocessed in Open Refine, an open source “data refinery” that was able to rid the data of inconsistent and incomplete values that would rather hurt the Subsequent stages of the said lifecycle.

On the analysis stage of the data Lifecycle, the Data was first uploaded into a Google Cloud bucket from where the following three technologies accessed it during the analytics stage: Hive, Spark and Big Query. Loading the data into the Hadoop file system was then achieved using a set of command line arguments. The process of moving the data is;

1. Local system to Hosted storage service in this case Google Cloud Bucket
2. From Google Cloud bucket to the DATA section of the manager Node
3. From the manager Node the data was then transferred into HDFS for processing using Hive and Spark.

Using simple SQL Queries, the data on Police Response to resistance was queried to identify trends, patterns and underlying characteristics in the data.

* Based on the simple analysis, the following aspects were discovered:  
  That districts, 2, 7 and 14 are the most notorious in Police resistance.
* That police officers with lower rank are prone to resistance
* That the Males are twice as much more likely to cause resistance during interactions with police as compared to women

Comparing the three tools Used during the analytics stage of this quest, it was discovered that While Spark boasted lighting quick speed, it was costly to implement due to high memory usage making it computationally inefficient. Hive on the other end was discovered to be simple to use even for beginners but had a problem with latency. This disqualifies it as a candidate for synthesizing a live data source. Google Big Query was commendable in performance and accessibility. It was fast to configure and easy to use as it has an interactive user interface.

The next two stages of the data lifecycle that need to be implemented on the data include Modelling using machine learning algorithms to understand deeper features and patterns that simple Explanatory data analysis can deliver. Another key step would be to create captivating visuals that pass the message to even non-technical audience.

**References**

Ali, M. H., Hosain, M. S., & Hossain, M. A. (2021). Big data analysis using BigQuery on cloud computing platform. *Australian Journal of Engineering and Innovative Technology*, 1–9. <https://doi.org/10.34104/ajeit.021.0109>

Mouad et. al., B. (2020). A new system for massive RDF data management using Big Data query languages Pig, Hive, and Spark. *International Journal of Computing and Digital Systems*, *9*(2), 259–270. <https://doi.org/10.12785/ijcds/090211>

Pryor, G. (2012). *Managing research data*. Facet Publishing.

Rahul, K., & Banyal, R. K. (2020). Data life cycle management in big data analytics. *Procedia Computer Science*, *173*, 364–371. <https://doi.org/10.1016/j.procs.2020.06.042>

Wijaya, A. (2022). *Data Engineering with Google Cloud Platform: A practical guide to operationalizing scalable data analytics systems on GCP*. Packt Publishing.