## Schedule Overview:

* Duration: April 25 to the end of May
* Timing: Every Tuesday and Thursday from 6:00 PM to 8:00 PM PST
* Session Length: 2 hours per class (24 hours)

## How will we pick topics:

We have more topics than time, so we will make sure to cover all topics in bold and we will pick up other topics based on students' needs.

## Training Modules:

1. **Introduction to Databricks & Delta**
   * Introduction to Databricks (PPT)
   * Delta Deep Dive (PPT)
   * dbdemos.install('delta-lake')
   * dbdemos.install('cdc-pipeline')
2. **Data Ingestion from No-SQL and SQL Sources**
   * Change Data Capture (CDC) Techniques
   * How to ingest from SQL
   * How to ingest from No-SQL
   * What is change data capture
   * Convert change data capture information into Delta
   * Integration with Kafka for Real-time Data Ingestion
   * Comparing Various Data Ingestion Products
   * dbdemos.install('cdc-pipeline')
   * dbdemos.install('dlt-cdc')
   * dbdemos.install('auto-loader')
   * Techniques for Batch Data Ingestion & Streaming
   * Handling Large-Scale Data with Efficiency
3. **Delta Live Tables - Databricks’ Flagship Product**
   * Introduction to Delta Live Tables
   * Building Real-time Data Pipelines
   * Advanced Features and Best Practices
   * dbdemos.install('dlt-cdc')
   * dbdemos.install('dlt-loans')
   * dbdemos.install('dlt-unit-test')
4. **Spark Streaming (2-3 Classes):**
   * Introduction to Spark Streaming and Kafka to Delta Streaming
   * Utilizing Auto Loader and understanding Delta operations (Append, Merge, Change Data Feed)
   * In-depth look at checkpoints in Spark Streaming
   * Advanced techniques: forEachBatch with Merge, Stream-Static Join, and Stream-Stream Join
   * dbdemos.install('streaming-sessionization')
   * [**https://canadiandataguy.com/**](https://canadiandataguy.com/)
5. Data Governance in Databricks
   * Managing Data Updates and Heavy Reads
   * Implementing Data Security and Compliance
   * Data Quality and Lineage in Databricks
6. Spark Tuning for Real-Life Scenarios
   * Understanding Spark Architecture
   * Read Spark white paper
   * Performance Tuning and Optimization
   * Troubleshooting Common Issues in Spark
7. Delta Deep Dive
   * Study of the Delta and Zorder White Papers
   * Understanding Liquid Clustering
   * Performance Tuning with Delta: Partitioning and Z-ordering
8. Choosing the Right Data Processing Product
   * Decision-Making Criteria for Batch vs. Stream Processing
   * Effective Use of Delta Live Tables
   * Scheduling and Managing Streaming Jobs
9. Advanced Spark Join Techniques
   * Optimizing with Broadcast Hash Join
   * Handling Data Skew in Joins
   * Best Practices for Efficient Spark Joins
10. **Unity Catalog Upgrade (2 Classes):**
    * Core concepts of Unity Catalog
    * Unity Catalog Migration Toolkit (UCX)
11. Optional Sessions:
    * Best practices on the Databricks platform
    * Introduction to the Well-Architected Framework

## Session Format:

* A mix of theoretical presentations and hands-on labs.
* Participants will use their **own Databricks environment preferably with Unity Catalog enabled**.

#### Materials:

* Comprehensive deck and additional resources will be provided for students to utilize during and post-training.
* <https://drive.google.com/drive/folders/1VXnsuAo0iUWqxOAlIdR_ubYvMnzhMslK?usp=sharing>