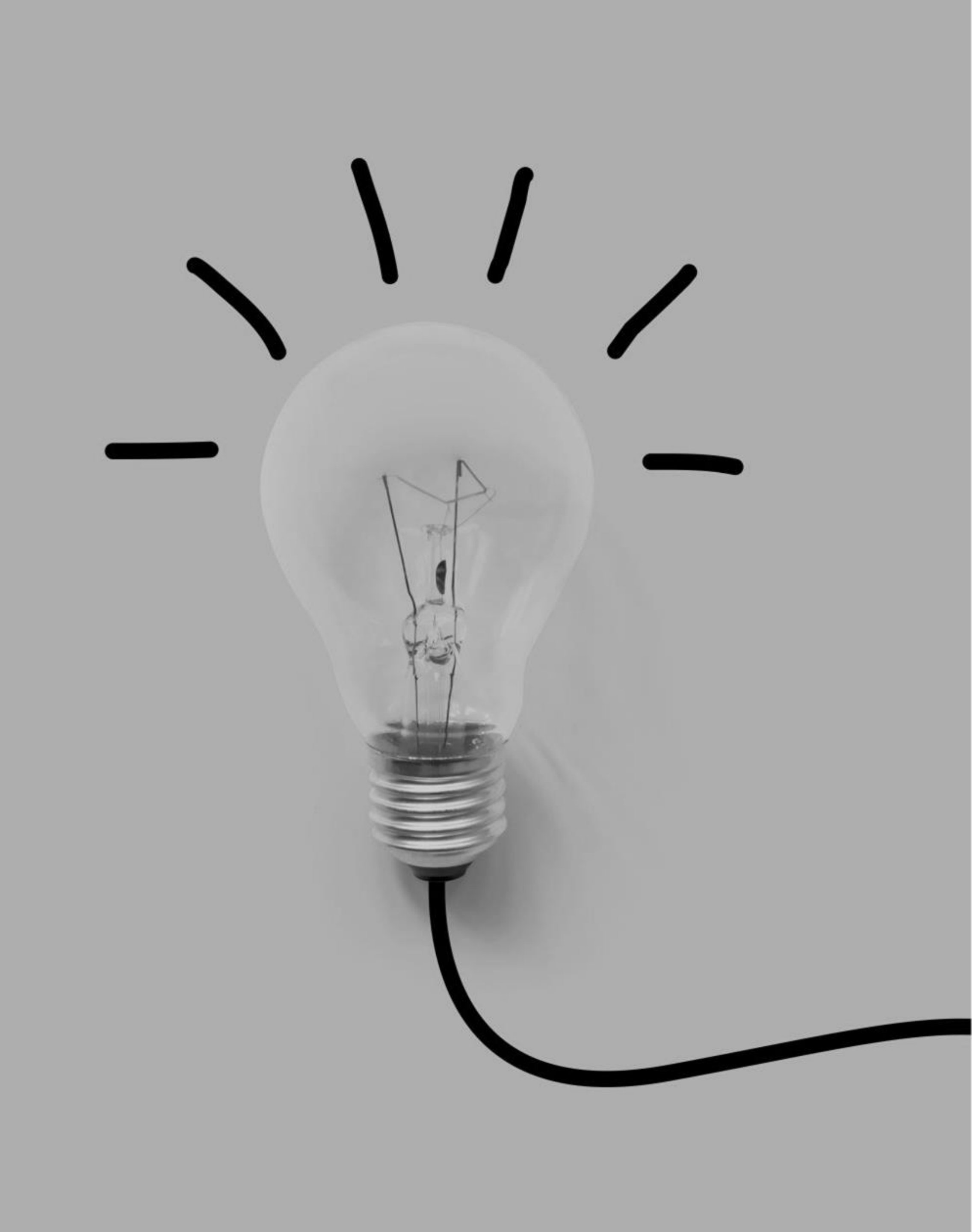
APACHE SPARK:
CASE STUDY

Chu Ngwoke



LEARNING OBJECTIVES

1. Introduction to Databricks
2. Spark DataFrames
3. Spark SQL
4. Data Cleaning in Spark
5. Data Transformation in Spark

10M PEOPLE DATASET

Imagine yourself as a diligent **data engineer**, employed by the **Central Bank** and entrusted with the critical task of overseeing the data integrity of commercial banks within the nation. One morning, as you settle into your office, a notification pops up on your screen—a dataset labeled “Spring Bank - Customers" has been dropped into the Central Bank's repository. **Your mission is clear: clean, analyze, extract insights from this data**, **aggregate the data, and load it into the Central Bank’s DWH for further analytics** to ensure compliance with regulatory standards and enabling informed decision-making by the banking authorities.

TASK 1: DATA EXPLORATION AND UNDERSTANDING

1. Explore the "10M-People" dataset to understand its structure and content.
2. Examine data types, potential data quality issues, and any sensitive information that requires anonymization.

TASK 2: DATA CLEANING AND TRANSFORMATION

3. Address missing or invalid values in the dataset.
4. Remove duplicate records to ensure data accuracy.
5. Convert all columns to appropriate data types.
6. Anonymize sensitive information such as social security numbers (SSN) to protect customer privacy.

TASK 3: AGGREGATION AND INSIGHT GENERATION

6. Aggregate the dataset to derive summary statistics, such as average salary by gender.
7. Generate insights into customer demographics, including gender distribution and age profiles.
8. Analyze salary distributions to understand income disparities among customers.

TASK 4: DWH LOADING

8. Design and create a Data Warehouse (DWH) table schema to store the cleaned and aggregated dataset.
9. Load the processed dataset into the DWH table for further analysis and reporting.

FIRE DEPARTMENT DATASET

Imagine yourself as a **dedicated data engineer**, employed by San Francisco **Fire Department** to ensure the efficiency and effectiveness of emergency response operations. One day, as you settle into your office at the Fire Department headquarters, you receive an urgent notification—an **extensive dataset containing records of calls made to the Fire Department** has been deposited into the department's data repository. **Your role is clear: clean, analyze, extract insights from this data, aggregate the data, and load it into the department's DWH for further analytics** to enhance emergency response strategies and optimize resource allocation for the benefit of the community.

TASK 1: DATA EXPLORATION AND UNDERSTANDING

1. Explore the “sf_fire_calls" dataset to understand its structure and content.
2. Examine data types and other potential data quality issues

TASK 2: DATA CLEANING AND TRANSFORMATION

3. Fix all the identified data quality issues

TASK 3: AGGREGATION AND INSIGHT GENERATION

4. How many distinct types of calls were made to the Fire Department?
5. What were distinct types of calls made to the Fire Department?
6. Find out all response for delayed times greater than 5 mins?
7. What were the most common call types?
8. What zip codes accounted for most common calls?

TASK 3: AGGREGATION AND INSIGHT GENERATION

9. How many distinct years of data is in the dataset?
10. What week of the year in 2018 had the most fire calls?
11. What neighborhoods in San Francisco had the worst response time in 2018?

TASK 4: DWH LOADING

12. Create an aggregate table to hold the total number of calls in each district in each month of each year
13. Design and create a Data Warehouse (DWH) table to store the aggregated dataset.
14. Load the aggregated data into the DWH table for further analysis and reporting.

THANK YOU

👤	☎	✉
Chu Ngwoke	-	chu.ngwoke@gmail.com