Big Data Processing Pipeline for Chicago Crime Data Analysis

Overview

This project implements a robust big data processing pipeline for real-time and historical analysis of the Chicago Crime dataset. By leveraging AWS cloud technologies, the pipeline efficiently ingests, processes, and visualizes crime data, offering actionable insights for urban safety, resource allocation, and policymaking.

Key Features

- **Real-Time Data Ingestion**: Ingests live data streams using AWS Kinesis.

- **Data Transformation**: Cleans and profiles data with AWS Glue.

- **Scalable Data Processing**: Analyzes data using Apache Spark on Amazon EMR.

- **Low-Latency Storage**: Stores processed data in DynamoDB and historical data in Amazon S3.

- **Interactive Dashboards**: Visualizes insights using Amazon QuickSight.

Architecture

The pipeline consists of the following stages:

1. **Data Ingestion**: AWS Kinesis streams for real-time ingestion.

2. **Data Transformation**: Data cleaning, profiling, and schema alignment using AWS Glue.

3. **Data Processing**: Spatial and temporal analysis via Apache Spark on Amazon EMR.

4. **Data Storage**: DynamoDB for real-time querying and Amazon S3 for historical data.

5. **Data Visualization**: Dashboards and heatmaps built in Amazon QuickSight.

Key Technologies

- **AWS Kinesis**: For high-throughput data ingestion.

- **AWS Glue**: For ETL (Extract, Transform, Load) operations.

- **Amazon EMR (Apache Spark)**: For large-scale data analysis.

- **AWS DynamoDB**: For low-latency querying.

- **Amazon S3**: For durable, scalable storage.

- **Amazon QuickSight**: For interactive data visualization.

Installation

1. **Clone the Repository**:

   ```bash
   git clone https://github.com/yourusername/your-repository-name.git
   cd your-repository-name
   ```

2. **Install Dependencies**:

   - Ensure you have Python and AWS CLI installed.

   - Install required Python libraries:

     ```bash
     pip install boto3 pandas pyspark
     ```

3. **Set Up AWS Credentials**:

   - Configure AWS CLI with your credentials:

     ```bash
     aws configure
     ```

4. **Create AWS Resources**:

   - Set up the necessary AWS resources, such as Kinesis streams, Glue jobs, EMR clusters, DynamoDB tables, and S3 buckets.

Usage

1. **Ingest Data**:

   - Run the Python script to ingest real-time or simulated crime data:

     ```bash
     python ingest_data.py
     ```

2. **Transform Data**:

   - Execute the AWS Glue job to clean and transform data.

3. **Process Data**:

   - Use Spark jobs on EMR for temporal and spatial analysis.

4. **Visualize Data**:

   - Access Amazon QuickSight dashboards for insights.

Example Visualizations

- Crime hotspots heatmap.

- Hourly and daily crime trends.

- Distribution of crime types by community areas.

Results

- **Spatial Insights**: Identified high-crime areas through clustering algorithms.

- **Temporal Insights**: Analyzed trends to detect peak crime hours and seasons.

- **Actionable Insights**: Dashboards providing intuitive visualizations for stakeholders.

Future Enhancements

- Integrate predictive analytics using AWS SageMaker.

- Enhance real-time processing with AWS Lambda.

- Add additional datasets, such as weather and demographic data.

Contributing

Contributions are welcome! Please follow these steps:

1. Fork this repository.

2. Create a feature branch:

   ```bash
   git checkout -b feature-name
   ```

3. Commit your changes and push:

```bash
git push origin feature-name
```

4. Create a pull request.

License

This project is licensed under the MIT License. See the LICENSE file for details.

Acknowledgments

- **Instructor**: Prof. Joseph Rosen

- **Team Members**: Pratiksha, Yash, Dinesh, Shivani

- **References**: [City of Chicago Open Data Portal](https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2)