EarthScape Climate Analytics System

Using Big Data, Hadoop, and Machine Learning

Introduction

Climate change is one of the most critical global challenges. Large-scale climate data generated from satellites, weather stations, and sensors requires advanced big data technologies for effective analysis and decision-making.

Project Requirements

The system must securely authenticate users, ingest climate data from multiple sources, store large datasets in a distributed environment, process batch and real-time data, apply machine learning models, visualize results, and generate alerts.

Functional Requirements

- User Authentication and Authorization
- Data Ingestion (Historical and Real-Time)
- Distributed Data Storage using HDFS
- Data Processing using Hadoop MapReduce
- Machine Learning for Prediction and Anomaly Detection
- Data Visualization through Dashboards
- Alerts and Notifications
- Feedback and Support System

Non-Functional Requirements

- High Performance and Optimization
- Data Security and Encryption
- Reliability with 99% Uptime
- Scalability and Load Balancing
- Compliance with Data Standards
- Complete Documentation

System Architecture and Flow

The system follows a layered architecture. Data is collected from climate sources, ingested into the system, stored in HDFS, processed using MapReduce, analyzed through machine learning models, visualized using dashboards, and monitored through alerts.

Technologies Used

Hadoop and HDFS for storage and processing, Apache Server for backend services, MongoDB and Impala for fast data access, Jupyter Notebook and RStudio for analytics, VS Code or PyCharm for development, and Tableau for visualization.

Conclusion

The EarthScape Climate Analytics System provides a scalable, secure, and intelligent platform for climate data analysis, supporting informed decisions to address climate change.

Architectural Flow

First, the user logs into the system. After authentication, data is ingested from various climate data sources. This data is stored in the Hadoop Distributed File System (HDFS). Hadoop MapReduce processes the data. Machine learning models are applied to the processed data. The results are visualized on dashboards, and if any anomaly is detected, alerts are generated.

Step 1: User Layer

- Web / Dashboard Interface
- Login & role verification

Step 2: Data Source Layer

- Satellite data
- Weather station data
- Sensor data

Step 3: Data Ingestion Layer

- Batch ingestion
- Real-time streaming

Step 4: Storage Layer

- HDFS (raw data)
- MongoDB / Impala (processed data)

Step 5: Processing Layer

- Hadoop MapReduce
- Data cleaning
- Pattern & anomaly detection

Step 6: Analytics & ML Layer

- Prediction models
- Trend analysis
- Model updates

Step 7: Visualization Layer

- Tableau dashboards
- Reports & charts

Step 8: Alert & Notification Layer

- Threshold checks
- Email / system alerts

Step 9: Monitoring & Security Layer

- Performance monitoring
- Access logs
- Encryption