Big Data Analysis of Clean Water & Sanitation (SDG 6)

Forecasting Access and Improving Water Quality

By Lazaro Martull

1. Problem Statement

Access to clean drinking water and safe sanitation remains one of the most significant global development and public health challenges. Over 2.2 billion people lack access to safely managed drinking water, and 3.6 billion people lack safe sanitation, as highlighted in global reports and reflected in our final presentation slides. These shortages contribute to widespread health issues, especially in low- and middle-income countries where waterborne diseases are prevalent and disproportionately impact young children. Poor water quality caused by pollution, untreated wastewater, and agricultural runoff further reduces the availability of safe water sources.

Communities without reliable access to clean water experience reduced economic productivity, higher healthcare costs, and time lost in water collection. The UN's Sustainable Development Goal 6 (SDG 6) aims for universal access to clean water and sanitation by 2030, but current trends indicate insufficient progress. Monitoring global water trends requires processing massive datasets, which are often inconsistent and too large for traditional systems.

To address this need, our project implemented a Big Data approach using Hadoop, HDFS, and MapReduce to process international water-quality data. We built a multi-node Hadoop cluster, developed Python-based mapper and reducer scripts, and used Hadoop Streaming to compute average water abstraction values by region and year. This workflow demonstrates how distributed systems can support large-scale analysis essential for understanding global water inequality.

2. Objectives

a. Forecast Progress Towards SDG 6:
    i. Analyze historical water-abstraction values to determine regional trends and disparities.
    ii. Use MapReduce output to identify which regions show improvement and which lag behind.
    iii. Provide scalable methods for assessing the gap between current progress and SDG 6 goals.
b. Analyze Water Quality:
    i. Use distributed computation to process and summarize water-quality indicators such as wat_basal_t.
    ii. Create a reproducible pipeline for large-scale water data analysis.

c. Identify Determinants:
    i. Categorize processed data by region and year.
    ii. Determine which regions display lower access or declining trends in safe water availability.
d. Provide Insights:
    i. Produce visualizations and summaries that highlight at-risk regions.
    ii. Offer data-driven observations that may inform future forecasting or policy recommendations.

3. Research Methodology

a. Our methodology focused on using Hadoop's distributed capabilities to process large water datasets. The original proposal included machine learning and forecasting plans; however, the final project centered on building and executing a complete MapReduce workflow due to the complexity and time required to configure a functional multi-node cluster.
b. Data Sources:
    i. WHO/UNICEF Joint Monitoring Programme (JMP) provided the structure and indicators used in the file.
    ii. World Bank water-access and quality indicators.
    iii. Supplementary data on regional classifications and measurement years.
    iv. Data was uploaded and stored in HDFS for distributed processing.
c. Data Workflow & Tools
    i. Hadoop HDFS + MapReduce:
        1. Used to ingest, filter, and compute aggregated statistics across large CSV files.
        2. Ensured parallelism, fault tolerance, and scalability.
    ii. Python Integration with Hadoop Streaming:
        1. Used Python scripts as mappers and reducers for key-value pair processing, cleaning, and averaging.
    iii. Cluster Setup:
        1. A three-node Hadoop cluster was built consisting of one master and two workers.
        2. Network configuration, XML configuration files, worker registration, and HDFS setup were performed manually.
        3. This setup is documented in our presentation slides showing the master and worker IP addresses, XML files, and job reports.

d. Research Process
    i. Preprocessing:
        1. The mapper handled header detection, index extraction, invalid-value skipping, and cleaning. It emitted output in the form: region,year value.
    ii. Exploratory Data Analysis:
        1. After MapReduce produced aggregated values, Python was used to generate line charts and bar graphs showing access trends and regional disparities.
    iii. Model Training and Validation:
        1. Not implemented due to final project constraints; focus was placed on completing the Hadoop processing pipeline.
    iv. Scalability Testing:
        1. Our MapReduce job successfully executed across all cluster nodes. Worker and master logs demonstrated distributed processing, block distribution, and job completion.
e. Expected Methodological Contributions
    i. A functioning Hadoop cluster capable of distributed processing.
    ii. A reproducible Python-based MapReduce workflow.
    iii. A data-cleaning and aggregation pipeline suitable for large-scale global datasets.
    iv. Summary statistics and visualizations identifying regional disparities in water access.
    v. A foundation for future forecasting, classification, and dashboarding work.

4. Research Activities

a. Literature Review
    i. Reviewed SDG 6 documentation, WHO/UNICEF JMP reports, and prior analyses on global water access.
b. Data Collection and Cleaning
    i. Uploaded raw CSV data to HDFS. The mapper script handled row-level cleaning.
c. Exploratory Data Analysis
    i. Created visualizations from reducer outputs to highlight regional differences and long-term trends.
d. Modeling and Forecasting
    i. Planned but not included in the final implementation; the completed focus was on distributed processing.

e. Model Evaluation and Validation
    i. Validated MapReduce summaries through sample manual calculations.

5. Time Schedule

Project Schedule	
Date	Task
Week 1-3	Literature review and define scope
Week 4-5	Collect and preprocess datasets in Hadoop
Week 6-7	Conduct exploratory data analysis
Week 8-9	Develop and train forecasting models
Week 9-12	Model evaluation and refinement
Week 13	Draft report and findings
Week 14	Final presentation and submission

6. Results and Findings

a. Regional Water Quality Summaries
    i. Using MapReduce, we successfully calculated average wat_basal_t values for each region and year. These outputs provide a clear overview of differences in water abstraction and highlight disparities that exist across global regions.
b. Identification of At-Risk Regions
    i. The reduced dataset and follow-up visualizations revealed that certain regions consistently show lower water-access metrics. These insights help identify areas that may require increased infrastructure investment or global support.
c. Big-Data Processing Workflow
    i. The project resulted in a functional Hadoop-based data processing pipeline that includes HDFS storage, a working multi-node cluster, Python mapper and reducer scripts, and successful execution through Hadoop Streaming. This pipeline is scalable and can be reused for much larger datasets.

d. Foundations for Future Work
    i. While forecasting models and dashboards were not implemented in the final project, the work completed provides a strong technical base for future teams to build predictive analysis, machine learning models, or interactive dashboards.

7. Conclusion

a. This project demonstrated how Hadoop and MapReduce can be used to process and analyze large-scale global water-quality data. By building a functional multi-node Hadoop cluster, developing Python-based mapper and reducer scripts, and executing the workflow using Hadoop Streaming, we were able to compute meaningful regional and yearly averages of wat_basal_t. The results provide insight into global disparities in clean water access and establish a reproducible foundation for future forecasting and modeling efforts.
b. Challenges and Limitations
    i. Configuring the Hadoop environment across multiple machines presented significant technical challenges, including networking issues, XML configuration mismatches, and path inconsistencies for Python execution. Due to the time required to stabilize the cluster, the forecasting and machine learning components originally planned were not implemented in this phase. Additionally, the dataset used for MapReduce processing contained missing or inconsistent values, requiring defensive cleaning logic in the mapper.

8. References

United Nations. (2023). *Sustainable Development Goal 6: Ensure availability and sustainable management of water and sanitation for all*. United Nations Department of Economic and Social Affairs. https://sdgs.un.org/goals/goal6

WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation and Hygiene. (2023). *Progress on household drinking water, sanitation and hygiene*. World Health Organization. https://washdata.org/

World Bank. (2023). *World Bank Open Data: Water and sanitation indicators*. The World Bank Group. https://data.worldbank.org/

Apache Software Foundation. (2023). *Hadoop Distributed File System (HDFS) architecture guide*. https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html

Apache Software Foundation. (2023). *MapReduce tutorial*. https://hadoop.apache.org/docs/stable/hadoop-mapreduce-client/hadoop-mapreduce-client-core/MapReduceTutorial.html

Apache Software Foundation. (2023). *Hadoop streaming*. https://hadoop.apache.org/docs/stable/hadoop-streaming/HadoopStreaming.html

World Health Organization. (2022). *Global analysis and assessment of sanitation and drinking-water (GLAAS)*. https://www.who.int/teams/water-sanitation-hygiene/glaas