**High-Level Design Document**

**Project:** GreenQuery – Sustainable BigQuery Optimization **Date:** July 22, 2025 **Team:** Lloyds Technology Centre

### 1. Introduction

GreenQuery is a sustainability-focused optimization platform designed to monitor, analyze, and improve BigQuery usage across GCP projects. It aims to reduce cloud costs, improve performance, and minimize environmental impact through intelligent monitoring, predictive modeling, and actionable recommendations.

### 2. Objectives

* Identify and reduce inefficient BigQuery queries
* Estimate and forecast cost and CO2 emissions
* Provide real-time and historical insights via dashboards
* Enable proactive alerting to data teams

### 3. Target Users

* Data Engineers
* Data Analysts
* Cloud FinOps Teams
* Engineering Leaders

### 4. Architecture Overview

**Core Components:**

* **BigQuery**: Data warehouse and central analytics engine
* **Log Sync**: GCP logging sink to central dataset
* **BQ Dataset**: Raw + processed audit log storage
* **BQ ML**: Forecasting and classification
* **Looker Studio**: Dashboard visualization
* **Users**: Interact through dashboards and alerts

### 5. Key Functional Modules

1. **Log Collector**
   * Uses GCP Log Sink to ingest BigQuery job logs
   * Stored in raw\_logs table
2. **Log Parser Function**
   * Extracts metadata and metrics (user, project, bytes, duration)
   * Writes to processed\_logs table
3. **Dry Run Estimator**
   * Executes dry\_run=True queries to simulate execution
   * Captures estimated bytes and cost
4. **Query Classification Engine (ML)**
   * Trains on historical logs using BigQuery ML
   * Labels queries: efficient vs inefficient
5. **Forecasting Engine (ML)**
   * Predicts future trends in cost and emissions
   * Supports ARIMA+ or regression models
6. **Optimization Advisor**
   * Static rule engine + ML-based suggestions
   * Examples: partition recommendation, filter addition
7. **Alert System**
   * Threshold-based alerts via Slack or Email
   * Configurable via YAML
8. **Dashboards**
   * Looker Studio powered visualizations
   * KPIs: Top inefficient queries, cost forecasts, emission heatmaps

### 6. Data Flow Summary

BigQuery Logs --> Log Sink --> Raw Dataset --> Log Parser --> Processed Logs  
 |  
 Dry Run Estimator  
 |  
 Query Classifier & Forecast Engine --> Insights Tables  
 |  
 Advisor --> Suggestions Table  
 |  
 Notifier --> Slack/Email  
 |  
 Looker Studio Dashboards

### 7. High-Level Technology Stack

| Layer | Technology |
| --- | --- |
| Cloud | Google Cloud Platform (GCP) |
| Compute | Cloud Functions, Scheduler |
| Storage | BigQuery (raw & processed datasets) |
| Monitoring | Cloud Logging, Ops Suite |
| ML | BigQuery ML (Classification & Forecasting) |
| Dashboards | Looker Studio |
| Alerting | Slack API, Email SMTP |

### 8. Non-Functional Goals

* **Cost Efficiency**: Leverage serverless functions and dry-run queries
* **Sustainability**: Forecast and reduce compute + emission impact
* **Security**: Least privilege IAM and log anonymization
* **Scalability**: Modular design allows growth across org projects

### 9. Success Metrics

* % reduction in inefficient query executions
* Estimated cost savings per month
* Reduction in CO2 emissions
* Time saved by engineering teams
* Alert response SLA compliance

### 10. Next Steps

* Complete Dry Run + ML modules
* Set up Log Sink and processed BQ datasets
* Build initial dashboards and notification templates
* Run pilot with internal data teams

**Prepared By:** GreenQuery Development Team Lloyds Technology Centre