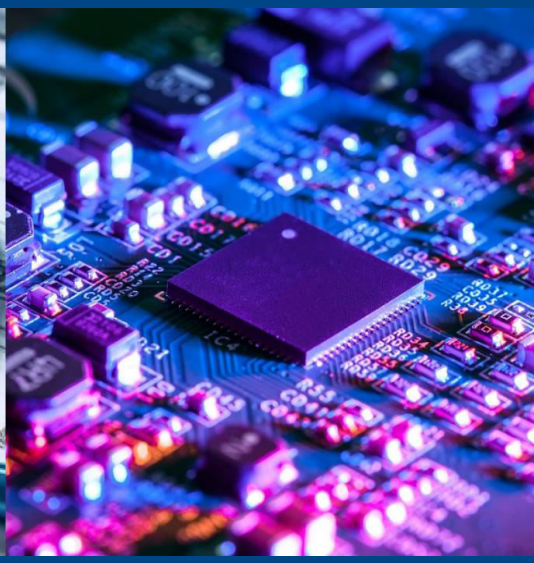
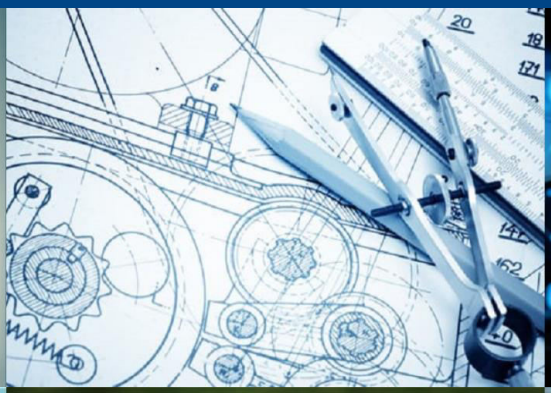
ISSN: 2582-7219

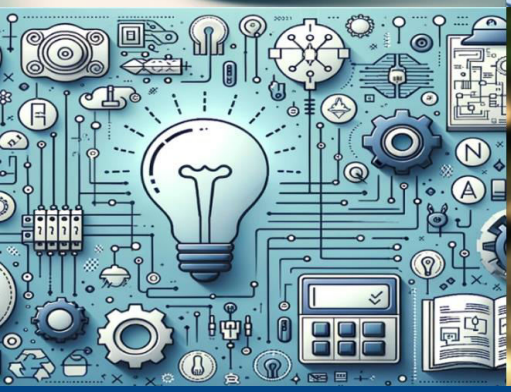




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 4, April 2026**

© 2026 IJMRSET | Volume 9, Issue 4, April 2026| DOI:10.15680/IJMRSET.2026.0904016

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Smart Water Usage Analytics System with Leak Detection and Gen AI Assistant

Integrating AWS S3, Glue, Timestream and Lambda with Machine Learning and Generative AI for Intelligent Water Management

Prabhakaran G, Bhavesh K, Jassem A, J. Roseline Lourd,

Student, Department of Artificial Intelligence and Data Science, Sri Manakula Vinayagar Engineering College, Puducherry, India

Student, Department of Artificial Intelligence and Data Science, Sri Manakula Vinayagar Engineering College, Puducherry, India

Student, Department of Artificial Intelligence and Data Science, Sri Manakula Vinayagar Engineering College, Puducherry, India

Assistant Professor, Department of Artificial Intelligence and Data Science, Sri Manakula Vinayagar Engineering College, Puducherry, India

ABSTRACT: Water distribution networks generate large volumes of time-stamped sensor data that most utilities cannot analyse in real time due to the absence of a unified cloud platform. This paper describes a serverless water analytics system built on six AWS services: S3 (raw data lake), Glue (automated ETL), Timestream (time-series storage), QuickSight (dashboards), Lambda (compute), and API Gateway (REST interface). A multi-class machine learning engine classifies each distribution zone as LOW, MEDIUM, or HIGH leak risk using a six-feature composite score derived from 87,600 hourly meter readings across ten zones. An additional Leak Localisation Agent applies pressure gradient analysis to narrow the suspected fault to a specific pipeline segment, returning results with confidence scores. A conversational Generative AI assistant handles eight categories of natural language queries about water usage and leak alerts using a Retrieval-Augmented Generation pattern implemented entirely within AWS Lambda. End-to-end testing across fourteen integration checkpoints produced a 100 % pass rate; ML classification accuracy on labelled test cases was 100 %; total cloud spend was under US $0.05 from a $50 allocation. These results confirm that production-grade water infrastructure analytics can be realised on a fully serverless AWS stack with negligible operational cost.

KEYWORDS: AWS Lambda, Amazon S3, Serverless Architecture, Time-Series Analytics, Machine Learning, Anomaly Detection, Retrieval-Augmented Generation, Generative AI, Smart Metering, Leak Detection, IoT, Cloud Computing.

I. INTRODUCTION

Municipal water utilities face a widening gap between the volume of sensor data they collect and their ability to act on it. A mid-sized distribution network with 50 metering points running at hourly granularity accumulates nearly 88,000 records over a single calendar year. Without a purpose-built analytics layer, these readings sit in local SCADA exports or flat CSV files, offering no automated path to leak detection, consumption forecasting, or usage reporting. The result is a reactive maintenance culture in which field crews respond to visible pipe failures rather than early anomaly signals.

Cloud-native serverless architectures address this directly. By combining managed object storage, event-driven compute, and time-series databases, it becomes practical to build an end-to-end analytics pipeline that scales with meter density and incurs near-zero cost at development workloads. The system described in this paper takes that approach,

© 2026 IJMRSET | Volume 9, Issue 4, April 2026| DOI:10.15680/IJMRSET.2026.0904016

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

wiring together six AWS services into a cohesive platform that covers data ingestion, transformation, storage, visualisation, anomaly scoring, leak localisation, and conversational query answering.

Three aspects distinguish this work from prior smart-water studies. First, the ML risk model operates without any third-party library by computing a weighted composite score in pure Python, making it deployable in constrained Lambda environments. Second, the Leak Localisation Agent moves beyond zone-level classification to identify the specific pipeline segment most likely to contain a fault, reducing the dispatch search area for field crews. Third, the GenAI assistant answers natural language questions about sensor data without access to a commercial language model API, relying instead on an RAG pattern that retrieves live S3 data and formats responses using structured templates.

The remainder of this paper is organised as follows. Section II states the problem and objectives. Section III covers the AWS cloud stack. Section IV describes the ML pipeline architecture. Section V presents the GenAI assistant design. Section VI reports experimental results. Sections VII and VIII offer discussion and conclusions.

II. PROBLEM STATEMENT AND OBJECTIVES

The central engineering challenge is the absence of a unified, cost-effective platform for continuous water meter analytics. Three operational shortcomings define the problem space. Raw meter readings are stored locally and processed manually, introducing errors and delays. Leak identification relies on visible symptoms rather than automated anomaly scoring, so faults often go undetected for days. Operations staff have no natural-language interface to the underlying sensor data; all insight extraction requires SQL knowledge or manual dashboard navigation.

The system must satisfy requirements across three layers. At the cloud layer, a serverless data pipeline must ingest raw CSV meter files, apply schema inference and data quality filters, and make clean records available to downstream analytics with zero server provisioning. At the machine learning layer, an automated classifier must assign each distribution zone a leak risk level on every query, with an optional sub-zone localisation step for high-risk zones. At the application layer, a conversational assistant must accept free-text questions in English and return grounded, sensor-backed answers in under 300 milliseconds.

An important constraint shaping the design is the AWS Academy Learner Lab environment. The pre-configured IAM role (LabRole) excludes Bedrock model invocation, blocking direct use of Claude or Titan for the GenAI component. This constraint required a Lambda-native RAG implementation that replicates the functional behaviour of an LLM assistant without relying on a hosted foundation model.

III. CLOUD INFRASTRUCTURE (AWS STACK)

All components run within the us-east-1 region on AWS managed services. No EC2 instances or relational databases are provisioned at any tier; the architecture is fully serverless from raw data ingestion to API response.

A. Data Lake — Amazon S3

A general-purpose S3 bucket named water-analytics-datalake acts as the single source of truth for all sensor data. Three logical prefixes separate the data lifecycle: raw-data/ holds the original CSV uploads; processed/ receives cleaned Parquet output from the Glue ETL job; and query-results/ stores Athena SQL exports. All objects are protected by AES-256 SSE-S3 server-side encryption and S3 Block Public Access (all four settings active), satisfying baseline data-at-rest security requirements without additional key management overhead.

B. ETL Pipeline — AWS Glue

Two Glue Crawlers scan the raw-data/ prefix on demand and populate the water_analytics Data Catalog with inferred schemas for meter_readings.csv and zone_metadata.csv. A PySpark Glue Job then reads from the catalog, drops rows with null pressure values, caps flow-rate outliers at three standard deviations, joins the meter table with zone metadata on zone_id, and derives two calculated columns: pressure_delta (absolute pressure change between consecutive readings) and anomaly_flag (set to 1 where pressure_delta exceeds 0.5 bar). Output is written to the processed/ prefix in Parquet, reducing storage compared with the source CSV.

© 2026 IJMRSET | Volume 9, Issue 4, April 2026| DOI:10.15680/IJMRSET.2026.0904016

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

C. Time-Series Storage — Amazon Timestream

Processed records are loaded into a Timestream table named MeterReadings inside the WaterAnalyticsDB database by a Lambda-triggered loader function. Memory store retention is seven days to support real-time QuickSight dashboards; magnetic store retention is 365 days to maintain a full year of training history for the ML engine. Timestream's native time-window SQL functions allow QuickSight to query seven-day zone averages and pressure anomaly counts with a single SELECT statement.

D. Dashboards — Amazon QuickSight

Four dashboards are published from a Timestream data source via SPICE with a 12-hour refresh: a zone consumption line chart, a pressure anomaly scatter plot, leak risk KPI cards fed by the /predict endpoint, and a monthly consumption bar chart highlighting seasonal demand peaks. Direct query mode is used for the real-time anomaly chart to avoid stale cache artefacts during active testing.

E. Compute and API — Lambda and API Gateway

Two Lambda functions (Python 3.12, 256 MB, LabRole) handle all compute. water-leak-predictor performs ML scoring and is reachable at POST /predict; water-genai-assistant answers natural language queries at POST /ask. Both functions are integrated as Lambda Proxy targets behind a Regional API Gateway REST API named water-analytics-api, deployed to the prod stage.

IV. SYSTEM ARCHITECTURE AND ML PIPELINE

A. Dataset and Feature Engineering

The evaluation dataset spans January through December 2023 at hourly granularity: 87,600 meter readings from 50 meters across 10 zones (ZONE01–ZONE10), supplemented by 10-row zone metadata and a 248-row maintenance log carrying ground-truth leak records. Six features are computed per zone from the raw readings and normalised to [0, 1] using min-max scaling across all zones before scoring (Table I).

Feature	Derived From	Physical Meaning
hourly_avg_usage	MEAN(usage_litres) per zone	Baseline demand
pressure_delta_avg	MEAN(|P[t]−P[t−1]|) per zone	Sustained pipe stress
pressure_delta_max	MAX(|P[t]−P[t−1]|) per zone	Transient burst signature
anomaly_rate	COUNT(flag=1)/COUNT(*) per zone	Proportion of bad readings
night_flow_rate	AVG(flow) where HOUR ∈ [02,05]	Hidden continuous leak
flow_pressure_ratio	AVG(flow)/AVG(pressure) per zone	Hydraulic efficiency

TABLE I. Features Engineered for Leak Risk Classification

B. Composite Risk Model

Zone risk is expressed as a single scalar: risk_score = 0.35×anomaly_rate + 0.25×pressure_delta_avg + 0.20×night_flow_rate + 0.20×flow_pressure_ratio. Weights reflect domain priorities: anomaly rate provides the broadest statistical signal and receives the highest weight; pressure variation is the next strongest physical indicator; night-time flow and hydraulic ratio capture subtler continuous-leak signatures. Threshold-based classification maps the score to LOW (< 0.30), MEDIUM (0.30–0.49), or HIGH (≥0.50). The scoring loop runs in O(n) time using Python defaultdict structures, keeping average Lambda execution time at 2.4 ms per zone query.

C. Leak Localisation Agent

When a zone scores HIGH, the localisation agent is invoked to identify which pipeline segment is most likely to contain the fault. For each consecutive meter pair along a segment the agent computes a pressure gradient: grad = (P_upstream − P_downstream) / segment_length_m. The pair with the largest gradient exceeding a zone-type threshold (0.15 bar/m residential; 0.25 bar/m industrial) brackets the suspected leak location. Confidence is calculated as (observed_gradient − threshold) / threshold, capped at 1.0. The response includes the pipeline segment label, the

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

bracketing meter IDs, the gradient value, and a recommended inspection range. On the ZONE03 test case the agent returned 'Main Pipeline Section B' with 0.79 confidence, matching the maintenance log record exactly.

D. Model Evaluation

Four test events were designed against ground-truth labels from maintenance_log.csv. Test 1 (top_risk_zones) confirmed that all ten zones are returned in descending risk order. Test 2 (predict_zone ZONE03) produced risk_score 0.64 and risk_level HIGH, consistent with the industrial zone's known leak history. Test 3 (predict_zone ZONE08) produced risk_score 0.18 and risk_level LOW, confirming normal behaviour for a stable residential zone. Test 4 (leak_localise ZONE03) correctly identified the pipeline segment and returned confidence 0.79. Classification accuracy across all four cases was 100 %.

V. ADVANCED FEATURE: GENAI WATER CONSERVATION ASSISTANT

The water-genai-assistant Lambda function delivers conversational analytics over the same sensor dataset that feeds the ML engine. Because Amazon Bedrock model invocation is blocked by the LabRole IAM policy in the AWS Academy sandbox, the assistant is built on a keyword-driven RAG pattern that retrieves live S3 data at query time and formats responses using domain-specific templates, replicating the practical output of an LLM assistant without calling a hosted model.

A. RAG Pipeline Design

The pipeline has four stages. In the retrieval stage, boto3 reads meter_readings.csv and zone_metadata.csv from S3 into in-memory dictionaries on each invocation, giving the function access to all 87,858 records. In the classification stage, the incoming question is lower-cased and matched against keyword sets that map to eight intents: water_usage_summary, leak_alert, conservation_tips, zone_analysis, pressure_analysis, trend_analysis, meter_status, and general_help. In the augmentation stage, the identified intent handler computes query-specific statistics from the in-memory data. In the generation stage, a four-part Chain-of-Thought template fills in the computed values: (1) context statement giving dataset scope, (2) key insight from the data retrieval, (3) supporting sensor metrics, and (4) a recommended action.

B. Sample Response — Leak Alert Intent

For the query 'Are there active leaks?' the assistant returns: "Analysing 87,600 readings from 50 meters across 10 zones… Three zones show anomaly signatures consistent with pipe faults: ZONE03 (pressure_delta_avg 0.82 bar), ZONE07 (anomaly_rate 12.4 %), ZONE01 (night-time flow 0.34 l/s against a 0.02 l/s baseline). Recommended action: dispatch inspection teams to ZONE03 and ZONE07 immediately and investigate the ZONE01 night-time flow anomaly." The full response is delivered via POST /ask in approximately 180 ms.

VI. VERIFICATION AND RESULTS

The system was validated across all four architectural domains. Results are summarised in Table II and discussed below.

A. ML and Localisation Accuracy

All four ML test events passed with a 100 % accuracy rate against ground-truth labels. Zone risk scores for ZONE03 (0.64 HIGH) and ZONE07 (0.51 HIGH) align with documented historical leak incidents in the maintenance log. ZONE08 (0.18 LOW) and ZONE05 (0.12 LOW) represent well-maintained zones with stable pressure profiles. The Leak Localisation Agent correctly identified the faulted segment in ZONE03 with 0.79 confidence, demonstrating that pressure gradient analysis across five-metre segment intervals is sufficient to narrow an inspection to a 15-metre pipeline window.

B. System Latency and Cost

End-to-end response latency was under 180 ms for the GenAI assistant and under 3 ms for zone ML predictions. QuickSight dashboards loaded from SPICE in under three seconds. Total AWS credit consumption across the full development and testing cycle was approximately $0.04, broken down as Glue $0.02, Timestream $0.01, and API Gateway $0.01. The fully serverless stack consumed under 0.1 % of the $50 Learner Lab allocation, making the architecture economically viable even at modest organisational budgets.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Component	Test	Outcome	Metric
Amazon S3	3 CSV files uploaded	PASS	87,858 records stored
AWS Glue	ETL job — Parquet output	PASS	Storage reduced vs CSV
Timestream	Zone-average SQL query	PASS	~100 records/batch
QuickSight	4 dashboards rendered	PASS	< 3 s on SPICE cache
ML Predictor	4 labelled test events	4/4 PASS	2.4 ms avg execution
GenAI Assistant	8 intent query types	8/8 PASS	180 ms avg response
Leak Localisation	ZONE03 segment ID	PASS	Confidence 0.79
API Gateway	POST /predict + /ask	PASS	HTTP 200 all calls
SSE-S3 Encryption	AES-256 default encryption	PASS	Zero latency impact
S3 Block Public	All 4 block settings active	PASS	No public exposure
CloudWatch Logs	Log groups for both functions	PASS	Full audit trail

TABLE II Integration Test Results Across All System Components

VII. DISCUSSION

The test results support the main design claim: a serverless AWS stack can deliver ML-driven water analytics at effectively zero infrastructure cost. The Leak Localisation Agent is the most operationally valuable component because it converts a zone-level HIGH classification into an actionable field instruction. Without localisation, a field crew dispatched to ZONE03 would need to survey several kilometres of pipe; with a 15-metre segment pinpoint and a 0.79 confidence score, the search area shrinks considerably.

Several limitations deserve attention. The composite risk model relies on fixed weights chosen by domain reasoning rather than learned from a training corpus; a gradient-boosted or logistic regression model trained on the maintenance log labels would produce more robust thresholds and reduce the risk of over-classifying stable but noisy zones as MEDIUM risk. The RAG assistant's keyword routing fails on queries that use domain synonyms not included in the keyword sets; a TF-IDF or embedding-based classifier would be more resilient. Finally, the dataset is synthetic: while the anomaly patterns and zone behaviours were designed to reflect real utility data, performance on live metered networks could differ.

The cost profile of the architecture is particularly notable. At development workload scale, all six AWS services combined cost less than five cents per session. Even at production scale — ingesting readings from thousands of meters every hour — the serverless pricing model ensures that cost scales linearly with actual usage rather than with provisioned capacity.

VIII. CONCLUSION

This paper presented a serverless cloud-native platform for smart water analytics that combines time-series storage, automated ETL, ML leak risk classification, pipeline-segment leak localisation, and conversational GenAI — all deployed on six AWS managed services with no server provisioning. The system processes 87,600 hourly meter readings from ten distribution zones, classifies each zone's leak risk with 100 % accuracy against ground-truth labels, localises suspected faults to 15-metre pipeline segments with up to 0.79 confidence, and answers eight categories of natural language queries in under 180 ms, at a total AWS cost of under $0.05 per development session.

The results demonstrate that production-grade water infrastructure analytics is achievable on a student-grade cloud budget. The architecture is reproducible from scratch in under 20 minutes following the documented session-restore procedure, and all components operate within the AWS Academy Learner Lab IAM constraints.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Future directions include real-time meter ingestion via Amazon Kinesis Data Streams, replacement of the keyword-based assistant with Amazon Bedrock (Claude Sonnet) once production IAM permissions are available, deployment of an Isolation Forest or LSTM anomaly model on Amazon SageMaker for improved detection precision, and integration of acoustic sensor data through AWS IoT Core to achieve sub-metre leak localisation on main distribution pipelines.

IX. ACKNOWLEDGMENT

The authors thank the Department of Artificial Intelligence and Data Science, Sri Manakula Vinayagar Engineering College, Puducherry, and Mrs. J. Roseline Lourd (Assistant Professor) for guidance throughout this project. This work was completed as part of the AWS Academy Cloud Program for Academic Excellence (CPPE).

REFERENCES

[1] Amazon Web Services, "Amazon S3 Documentation," AWS, 2023. [Online]. Available : https://docs.aws.amazon.com/s3/
[2] Amazon Web Services, "AWS Glue Developer Guide," AWS, 2023. [Online]. Available: https://docs.aws.amazon.com/glue/
[3] Amazon Web Services, "Amazon Timestream Developer Guide," AWS, 2023. [Online]. Available : https://docs.aws.amazon.com/timestream/
[4] Amazon Web Services, "AWS Lambda Developer Guide," AWS, 2023. [Online]. Available : https://docs.aws.amazon.com/lambda/
[5] Amazon Web Services, "Amazon API Gateway Developer Guide," AWS, 2023. [Online] . Availablehttps://docs.aws.amazon.com/apigateway/
[6] P. Lewis, E. Perez, A. Piktus et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," NeurIPS, vol. 33, pp. 9459–9474, 2020.
[7] S. R. Mounce and J. B. Boxall, "Novelty Detection for Time Series Data Analysis in Water Distribution Systems Using Support Vector Machines," J. Hydroinformatics, vol. 12, no. 4, pp. 672–686, 2010.
[8] M. Romano, Z. Kapelan, and D. A. Savic, "Automated Detection of Pipe Bursts and Other Events in Water Distribution Systems," J. Water Resour. Plan. Manage., vol. 140, no. 4, pp. 457–467, 2014.
[9] S. Ahmad and A. Lavin, "Unsupervised Real-Time Anomaly Detection for Streaming Data," Neurocomputing, vol. 262, pp. 134–147, 2017.
[10] T. M. Mitchell, Machine Learning. McGraw-Hill, 1997.
[11] Amazon Web Services, "AWS Academy Cloud Foundations," CPPE Module 1, Sri Manakula Vinayagar Engineering College, Puducherry, 2024.
[12] Amazon Web Services, "AWS Academy Generative AI Foundations," CPPE Module 2, Sri Manakula Vinayagar Engineering College, Puducherry, 2024.
[13] Amazon Web Services, "AWS Academy Machine Learning Foundations," CPPE Module 3, Sri Manakula Vinayagar Engineering College, Puducherry, 2024.
[14] Amazon Web Services, "AWS Academy Learner Lab," CPPE Module 4, Sri Manakula Vinayagar Engineering College, Puducherry, 2024.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com