Cloud Cost Optimization & Anomaly Detection

Project title: Cloud Cost Optimization & Anomaly Using Data Analytics

Date: 18th September, 2025

Background & Business Objective

With enterprises rapidly scaling their operations on cloud platforms (AWS, Azure, GCP), cloud bills have become a major controllable cost factor. Many IT companies face:

- Unexpected spikes in usage leading to surprise bills.
- Underutilized or idle resources consuming budget.
- Anomalies caused by misconfigurations, attacks, or runaway jobs.

Business leaders (CFOs, CTOs, Engineering Managers) need data-driven visibility into usage and spend patterns to forecast budgets accurately, detect anomalies early, and enforce governance.

This project's objective is to demonstrate how data analytics can reduce waste, improve efficiency, and prevent unexpected cloud overspend.

Business Problem

Cloud cost overruns and anomalies lead to:

- 20–40% overspending annually due to lack of governance and transparency.
- Difficulty identifying which team/service/region caused the spike.
- Inability to predict future spend accurately, leading to budget shocks.
- SLA breaches when anomalies impact customer-facing workloads.

Without robust anomaly detection and forecasting, IT companies lose millions and undermine operational efficiency.

Goal

To build an end-to-end analytics solution that:

- Detects and flags anomalous cloud spending in near real-time.
- Forecasts future spend using statistical and machine learning models.
- Identifies cost-saving opportunities by analyzing underutilized resources.
- Provides actionable dashboards for both executives (strategic view) and engineers (operational root-cause analysis).

Scope of Work

1. **Dataset Creation:**

- Generate a synthetic dataset with 1M+ rows of cloud usage & billing logs.
- Include realistic messiness (missing values, duplicates, outliers).

2. **Data Analytics (Python Jupyter):**

- Data cleaning, preprocessing, and feature engineering.
- Descriptive statistical testing (patterns in spend).
- Time-series forecasting (ARIMA/SARIMAX/Prophet).
- Anomaly detection (Isolation Forest, rolling z-scores).

3. **Insights & Business Value:**

- Root cause analysis (which service/team caused spikes).
- Potential cost savings through optimization.
- Impact of anomalies on financial forecasts.

4. **Visualization (Power BI):**

- KPI dashboards with filters (date, service, team, region).
- Trend analysis with anomaly overlays.
- Drill-down into service/team-level costs.
- What-if analysis: potential savings if anomalies were prevented.

Tools & Technologies

- Python (Jupyter Notebooks): pandas, NumPy, matplotlib, seaborn, scikit-learn, statsmodels, Prophet.
- Data Storage: CSV, Parquet for optimized processing.
- Machine Learning: Isolation Forest, Local Outlier Factor, Prophet.
- Visualization: Microsoft Power BI (interactive dashboards with KPIs, slicers, drilldowns).
- Version Control: GitHub (for portfolio presentation).

Deliverables

1. Synthetic Dataset (CSV, ≥1M rows) simulating real-world cloud usage & cost logs.
2. Cleaned & Processed Dataset (Parquet/CSV) ready for analytics.
3. Jupyter Notebooks:

- Data cleaning & preprocessing.
- Statistical & ML-based anomaly detection.
- Time-series forecasting & insights.

4. Power BI Dashboard (.pbix):

- Page 1: Executive KPIs & trends.
- Page 2: Root-cause analysis drilldown.

- Page 3: Model performance & what-if savings.

5. Executive Report (README/PDF):

- Key insights, estimated cost savings, recommendations.