Building a Scalable Data Pipeline for Airbnb Listings Analytics in NYC

[Airbnb Listings Data Pipeline- Business-Focused Data Pipeline for Hospitality Insights]

[Bootcamp - AWS Data Engineering Project - 5]

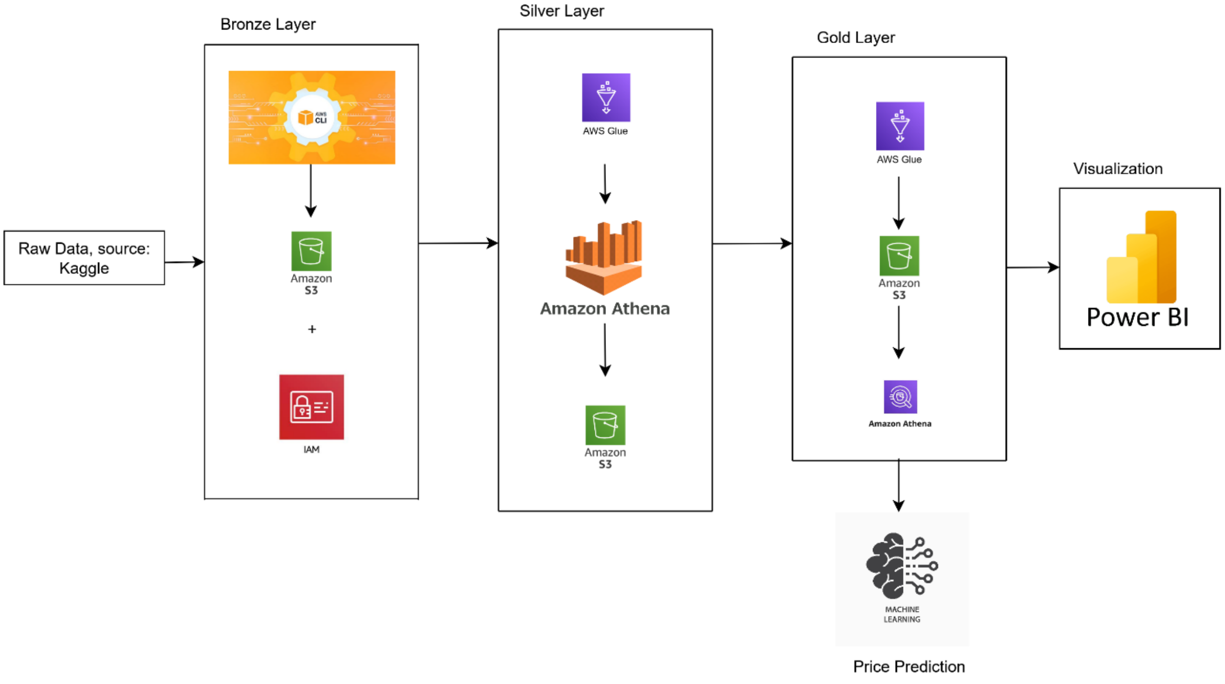
Objective

The objective of this project is to design and implement a scalable data pipeline that transforms raw Airbnb listing data from New York City into actionable business insights. The pipeline aims to:

- Support revenue optimization through price benchmarking and anomaly detection.
- Enhance operational efficiency by automating host performance tracking.
- Enable market analysis by identifying neighborhood-level demand patterns.
- Improve customer experience by monitoring review trends and availability.

The solution leverages AWS services to ensure scalability, cost-efficiency, and business agility while handling geographic and temporal data.

System Architecture



Prerequisites

To implement this pipeline, the following prerequisites are required:

- AWS Account with permissions for S3, Glue, and IAM.
  - ✓ Source Data: Airbnb NYC dataset (airbnb_nyc.csv).
- AWS CLI for initial data uploads.
- Python/PySpark for Glue job scripting.
- BI Tool (Power BI) for visualization.

Component Breakdown

Component	Purpose
Amazon S3	Stores raw (Bronze), cleaned (Silver), and business-ready (Gold) data.
AWS Glue	ETL processing, data cataloging, and schema enforcement.
AWS Glue Crawlers	Automatically infers schema from S3 data.
Amazon Athena	Query Silver/Gold layer data using SQL.
Power BI	Visualization for business insights.
AWS IAM	Securely grants permissions to Glue, S3, and other services.

Design Decisions

1. Multi-Layered Architecture
   - Ensures data quality progression (raw -> cleaned -> business-ready).
   - Enables reprocessing flexibility without losing raw data.
2. AWS Glue for ETL
   - Serverless, scalable, and integrates with Glue Data Catalog for metadata.
3. Athena over Redshift
   - Used for ad-hoc queries to avoid Redshift costs for small-scale analytics.

Data Flow

1. Ingestion (Bronze)
   - Raw data uploaded via CLI (aws s3 cp).
   - Stored in s3://neha-bc005-airbnb-pipeline/bronze/.
2. Processing (Silver)
   - Glue Crawler infers schema.
   - Glue Job cleans, enriches, and stores data in s3://neha-bc005-airbnb-pipeline/silver/.
3. Business Layer (Gold)

- Glue Job transforms data into star schema (dim_listings, fact_reviews).
- Stored in s3://neha-bc005-airbnb-pipeline/gold/.

4. Analytics
   - Athena queries for insights (e.g., SELECT neighbourhood_group, AVG(price)).
   - Power BI visualizes trends (e.g., price by room type).

Security & Compliance

1. Encryption:
   - S3 Server-Side Encryption (SSE-KMS) for data at rest.
   - TLS for data in transit.
2. Access Control:
   - IAM Roles restrict Glue/S3 permissions.
   - Lake Formation enforces neighborhood-based data access.

Monitoring & Quality

1. Pipeline Health:
   - CloudWatch monitors Glue job failures.
2. Data Quality:
   - Automated checks on critical fields (price, availability).
   - Glue Data Quality rules (future enhancement).
3. Cost Optimization:
   - S3 Lifecycle Policies to archive old data to Glacier.
   - Glue G.1X workers for cost-efficient processing.