

Visualizing Housing Market Trends: An Analysis of Sale Prices and Features

1. High-Level Architecture Overview

Data Sources → Data Ingestion → Data Storage → Data Processing → Analytics Engine → Visualization Layer → End Users

2. Architecture Components

A. Data Sources:

- Real estate transaction databases
- Historical MLS data
- CRM systems
- Public property records
- Government housing datasets
- APIs (Zillow, Redfin)

B. Data Ingestion Layer:

- Batch ETL pipelines
- Streaming via Kafka
- Tools: Airflow, AWS Glue, Azure Data Factory

C. Data Storage Layer:

- Data Lake (S3, Azure Data Lake)
- Data Warehouse (Snowflake, Redshift, BigQuery)
- Star Schema: Fact (Sales), Dimensions (Property, Location, Time)

D. Data Processing:

- Cleaning & Outlier Removal
- Feature Engineering (Price per sqft, YoY growth)
- Tools: PySpark, Pandas, dbt

E. Analytics & Modeling:

- Descriptive Analytics
- Diagnostic Analytics
- Predictive Models (Regression, XGBoost, ARIMA)

F. Visualization Layer:

- BI Tools (Tableau, Power BI, Looker)
- Dashboards: Market Trends, Feature Impact, Geographic Trends, Forecast

G. End Users:

- Analysts, Investors, Developers, Buyers, Executives

3. Data Flow

APIs/CSV → Ingestion → Data Lake → ETL → Data Warehouse → Analytics → BI Dashboards

4. Scalability & Performance

- Partition by date and region
 - Columnar storage (Parquet)
 - Materialized views
 - BI caching
-

5. Security & Governance

- Role-based access control
 - Data masking
 - Encryption at rest and transit
 - Audit logging
-

6. Deployment

- Cloud: AWS, Azure, GCP
 - CI/CD: GitHub Actions
 - Infrastructure as Code: Terraform
-

7. KPIs

- Median sale price
 - Price per sqft
 - Inventory rate
 - Time on market
 - YoY growth
 - Forecast accuracy
-