Solution Architecture: Leading Real Estate Company (India) on Data Analytics

Company Overview

A leading real estate company in India operates across multiple cities, managing residential, commercial, and industrial properties. The company aims to optimize data integration, improve analytics capabilities, and enhance reporting to support decision-making.

Business Challenges

1. **Diverse Data Sources**:
    - On-premises legacy databases store property and client information.
    - Excel files contain ad-hoc data from brokers and agents.
    - Cloud-based systems manage CRM and marketing data.
2. **Siloed Data**:
    - Disconnected data makes it challenging to gain a unified view of property performance, sales trends, and customer behavior.
3. **Performance Bottlenecks**:
    - Inefficient data transformation processes lead to delayed insights.
    - High reporting latency impacts business decisions.
4. **Scalability Issues**:
    - Increasing data volume requires scalable infrastructure to handle future growth.

Solution Overview

The company implemented a **data analytics solution** leveraging Apache Spark, a cloud-based Data Warehouse, and Power BI for reporting.

Raw Data Sources:

- **On-premises Database**: Property sales data with 10 million rows.
- **Excel Sheets**: Agent performance data with 500,000 rows.
- **Cloud CRM**: Customer data with 2 million rows.

Data Flow and Architecture

1. **Data Ingestion**:
    - On-premises data ingested using Spark's JDBC connector.
    - Excel files uploaded to Azure Blob Storage.
    - Cloud-based CRM and marketing data fetched via REST APIs.
2. **Data Lake Architecture**:
    - **Bronze Layer**: Raw data from all sources is stored in Azure Data Lake (Parquet format).

On-premises Sales Data:

- Sales data is fetched from a SQL Server database using Spark JDBC.
- **Data size**: 10 million rows (~5 GB).

Excel Data:

- Broker performance data uploaded to Azure Blob Storage (500,000 rows, ~100 MB).

Cloud CRM Data:

- Customer data (2 million rows, ~1 GB) fetched via API and saved in JSON format.

- **Silver Layer**: Spark cleans and standardizes the data. Key operations include:
  - Deduplication of customer records.
  - Formatting inconsistent date fields.
  - Standardizing currency formats.

  Deduplication and Cleaning:

  - Remove duplicate sales records.
    Before Deduplication: 10 million rows.
    After Deduplication: 9.8 million rows (~2% duplicates).

  Join Datasets:

  Join sales data with broker and customer data to enrich it with additional details.

  - **Sales Data**: 9.8 million rows.
  - **Broker Data**: 500,000 rows.
  - **Customer Data**: 2 million rows.

- **Gold Layer**: Spark aggregates and enriches data for analytical queries. Examples:
  - Monthly property sales summaries.
  - Customer segmentation for targeted marketing.

  Aggregations for Gold Layer:

  - Calculate monthly sales revenue by city and property type:
    Input: 9.8 million rows.
    Output: 100 rows (monthly aggregates).

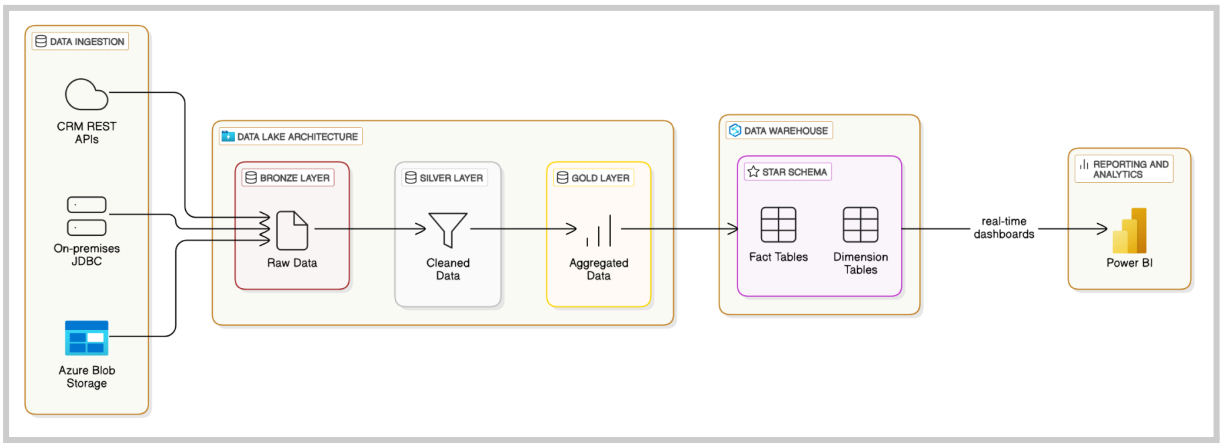
3. **Data Warehouse**:
   - Optimized data from the Gold Layer is loaded into **Azure Synapse Analytics** using Spark.
   - Schema: Star schema with fact tables (e.g., Sales, Revenue) and dimension tables (e.g., Properties, Customers).

- **Monthly Aggregates**: 100 rows (~200 KB).
- **Enriched Transactional Data**: 9.8 million rows (~8 GB).

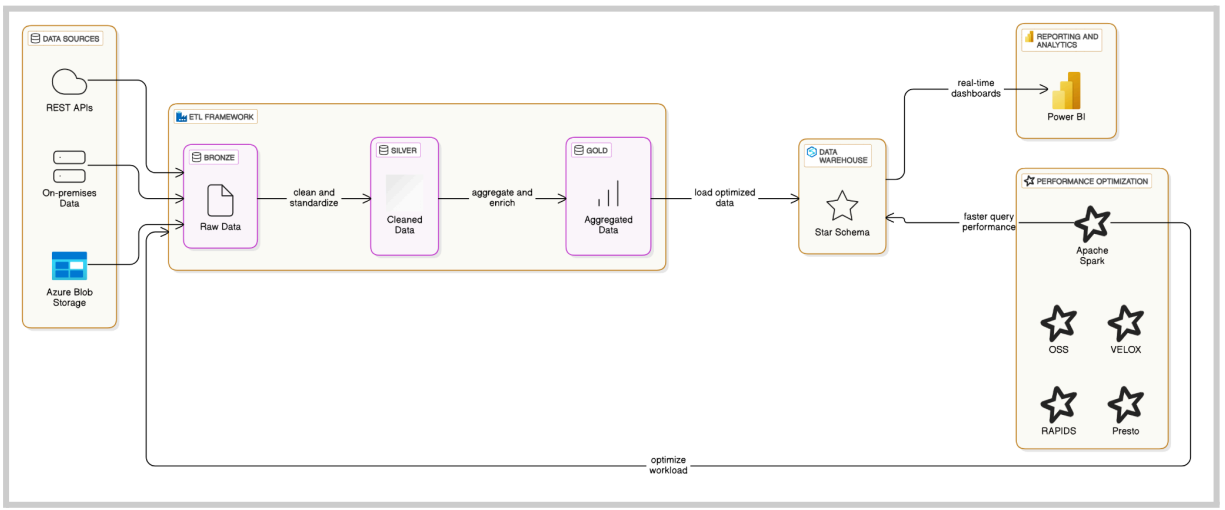
4. **Reporting and Analytics**:
   - Power BI connects directly to the Data Warehouse, enabling real-time dashboards for:
     - Sales performance by city and property type.
     - Broker efficiency and contribution.
     - Marketing campaign ROI.

     - **Monthly Sales Dashboard**:
       - Sales revenue by city: **10 cities, 12 months → 120 data points.**
     - **Top Brokers Dashboard**:
       - Broker performance metrics based on **500,000 rows**.
     - **Customer Segmentation Dashboard**:
       - Analyze behavior across 2 million customer records.

Architecture Diagram: 1



Architecture Diagram: 2 (Performance Optimization with **Apache Spark**)



Results Achieved

1. **Unified Data Platform**:
    - Consolidated data from on-premises, Excel, and cloud systems into a single source of truth.
2. **Improved Performance**:
    - Spark's optimizations reduced ETL job runtimes by 40%.
3. **Enhanced Decision-Making**:
    - Real-time dashboards in Power BI enabled faster and more informed decisions.
4. **Scalability**:
    - The solution scales seamlessly to accommodate increasing data volume and additional data sources.
5. **Operational Efficiency**:
    - Automated data pipelines reduced manual effort in handling ad-hoc requests.

Image: Real Estate Sales Power BI Dashboard