Laporan Proyek Pipeline Data untuk Prediksi Harga Akomodasi Pariwisata di Yogyakarta

Topik

Proyek Infrastruktur Pipeline Data untuk Prediksi Harga Akomodasi Pariwisata di Yogyakarta

Permasalahan

Pasar pariwisata di Daerah Istimewa Yogyakarta (DIY) mengalami dinamika yang kompleks akibat fluktuasi harga, kebijakan, dan faktor ekonomi. Perubahan harga akibat musim wisata, biaya operasional, dan kompetisi antar akomodasi sering kali menyulitkan pelaku bisnis dalam menentukan harga yang optimal. Selain itu, preferensi wisatawan yang beragam juga menambah tantangan dalam pengelolaan harga akomodasi.

Analisis data pasar menjadi solusi untuk memahami tren harga dan memprediksi dinamika pasar pariwisata. Dengan menggabungkan data hotel dari multiple platform, faktor geospasial, amenitas, dan reputasi online, pelaku usaha akomodasi dapat mengoptimalkan strategi harga, meningkatkan okupansi, dan memaksimalkan keuntungan secara efektif.

Analisis prediktif harga pariwisata di DIY memerlukan sumber data yang mencakup informasi tentang:

- **Faktor Geospasial**: Kedekatan dengan pusat transit dan objek wisata
- **Amenities**: Fasilitas yang ditawarkan hotel seperti Wi-Fi gratis dan sarapan
- **Reputasi Online**: Skor ulasan dan visibilitas platform pemesanan
- **Cross-Platform Pricing**: Perbandingan harga antar platform booking

Data Yang Digunakan

Berdasarkan hasil scraping yang telah dilakukan, data yang tersedia meliputi:

1. Data hasil scraping informasi akomodasi dari Booking.com
2. Data hasil scraping ulasan akomodasi dari Booking.com
3. Data hasil scraping informasi akomodasi dari TripAdvisor
4. Data hasil scraping ulasan akomodasi dari TripAdvisor
5. Data geospatial lokasi objek wisata dan tempat makan di DIY

Deskripsi Raw Data dari Data Resource

1. Data Hasil Scraping Informasi Akomodasi dari Booking.com

Sumber: Hasil scraping dari situs Booking.com untuk akomodasi di DIY.

Struktur data yang tersedia:

- **Informasi dasar**: `name`, `type`, `description`, `stars`, `hotelId`
- **Lokasi**: `location.lat`, `location.lng`, `address.full`, `address.street`, `address.country`, `address.region`, `address.postalCode`
- **Harga dan mata uang**: `price`, `currency`
- **Rating dan ulasan**: `rating`, `reviews`
- **Fasilitas**: Array `facilities` dengan struktur nested berisi kategori fasilitas dan detail:
  - Kamar Mandi (Tisu toilet, Handuk, Sandal, dll)
  - Kamar Tidur (Seprai, Lemari)
  - Media/Teknologi (TV layar datar, TV satelit, Telepon)
  - Internet (Wi-Fi gratis di seluruh hotel)
  - Tempat Parkir (Parkir pribadi gratis)
  - Layanan (Resepsionis 24 jam, Layanan kebersihan harian)
- **Informasi operasional**: `breakfast`, `checkIn`, `checkOut`
- **URL**: Link ke halaman hotel di Booking.com

2. Data Hasil Scraping Ulasan Akomodasi dari Booking.com

Sumber: Hasil scraping dari situs Booking.com untuk ulasan akomodasi di DIY.

Struktur data yang tersedia:

- **Identitas**: `id`, `hotelId`, `stayRoomId`
- **Ulasan**: `rating`, `likedText`, `dislikedText`, `reviewTitle`
- **Rating kategori**: Array `hotelRatingScores` dengan kategori:
  - Staff, Facilities, Cleanliness, Comfort, Value for money, Location
- **Info pengguna**: `travelerType`, `userLocation`, `userName`
- **Waktu**: `checkInDate`, `checkOutDate`, `reviewDate`, `numberOfNights`
- **Detail kamar**: `roomInfo`
- **Metadata**: `helpfulVotes`, `totalCategoryReviews`, `reviewLanguage`

3. Data Hasil Scraping Informasi Akomodasi dari TripAdvisor

Sumber: Hasil scraping dari situs TripAdvisor untuk akomodasi di DIY.

Struktur data yang tersedia:

- **Informasi dasar**: `name`, `description`, `type`, `category`, `subcategories`
- **Lokasi**: `latitude`, `longitude`, `address`, `addressObj`
- **Rating**: `rating`, `rawRanking`, `ratingHistogram`, `numberOfReviews`
- **Peringkat**: `rankingPosition`, `rankingDenominator`, `rankingString`
- **Fasilitas**: Array `amenities` dengan 25+ fasilitas berbeda
- **Harga**: `priceLevel`, `priceRange`
- **Kontak**: `phone`, `email`, `website`
- **Foto**: `photoCount`, `image`
- **Tips kamar**: Array `roomTips` dengan saran dari tamu
- **Info tambahan**: `hotelClass`, `localName`, `checkInDate`, `checkOutDate`

4. Data Hasil Scraping Ulasan Akomodasi dari TripAdvisor

Sumber: Hasil scraping dari situs TripAdvisor untuk ulasan akomodasi di DIY.

Struktur data yang tersedia:

- **Identitas**: `id`, `locationId`
- **Ulasan**: `rating`, `text`, `title`, `lang`
- **Waktu**: `publishedDate`, `travelDate`
- **Info pengguna**: Nested object `user` dengan `userId`, `name`, `contributions`, `userLocation`
- **Trip info**: `tripType` (FAMILY, SOLO, dll)
- **Foto**: Array `photos` dengan detail gambar ulasan
- **Respon pemilik**: `ownerResponse`
- **Info properti**: Nested object `placeInfo` dengan detail hotel

5. Data Geospatial Lokasi Objek Wisata di DIY

Sumber: Hasil scraping dari Google Maps untuk objek wisata dan tempat makan di DIY.

Struktur data yang tersedia:

- **Informasi dasar**: `title`, `categoryName`, `categories`

- **Lokasi**: `location.lat`, `location.lng`, `address`, `neighborhood`, `city`, `state`, `postalCode`
- **Kontak**: `phone`, `phoneUnformatted`
- **Rating**: `totalScore`, `reviewsCount`, `imagesCount`
- **Operasional**: Array `openingHours` dengan jadwal buka per hari
- **Status**: `permanentlyClosed`, `temporarilyClosed`
- **Fasilitas detail**: Nested object `additionalInfo` dengan kategori:
  - Service options (Delivery, Takeout, Dine-in)
  - Accessibility (Wheelchair accessible)
  - Offerings (Halal food, Quick bite)
  - Amenities (Restroom)
  - Atmosphere (Casual, Cozy, Trendy)
  - Payments (Credit cards)
  - Parking (Free parking lot)
- **Identifiers**: `placeId`, `fid`, `cid`, `kgmid`

Transformasi Data

Data Cleansing

- Menghapus/mengubah data duplikat dan nilai missing
- Standarisasi format tanggal, angka, mata uang (IDR normalization)
- Menghilangkan outlier ekstrem pada rating dan harga
- Memperbaiki format teks pada ulasan (menghapus karakter khusus, emoji)
- Standardisasi nama hotel untuk matching antar platform
- Normalisasi struktur nested facilities dan additionalInfo

Data Integration & Matching

- Hotel matching berdasarkan nama hotel dan koordinat geospasial antara TripAdvisor dan Booking.com
- Identifikasi hotel duplikat menggunakan similarity nama dan jarak koordinat (<100m)
- Merge data hotel yang sama dari kedua platform
- Cross-platform price comparison untuk hotel yang tersedia di kedua platform
- Platform source labeling untuk tracking asal data

Data Structuring

- Mengelompokkan akomodasi berdasarkan jenis (`type`, `category`), bintang (`stars`, `hotelClass`), dan lokasi
- Menstandarisasi rating scale (TripAdvisor 1-5, Booking.com conversion)
- Harmonisasi kategori fasilitas antar platform (mapping WiFi, breakfast, parking)
- Standardisasi rating categories (Staff, Cleanliness, Location, Value, dll)
- Flattening nested structures untuk analisis yang lebih mudah

Data Enriching

- Menghitung jarak akomodasi dari objek wisata menggunakan Haversine formula
- Membuat indeks aksesibilitas berdasarkan additionalInfo.Accessibility
- Menambahkan kolom density objek wisata dalam radius 1km, 2km, 5km
- Mengembangkan skor popularitas berdasarkan reviewsCount dan totalScore
- Menghitung rata-rata rating berdasarkan kategori dari hotelRatingScores
- Ekstraksi sentimen dari text, likedText, dislikedText
- Price differential analysis antar platform
- Availability scoring berdasarkan ketersediaan data di berbagai platform

Struktur Data Hasil Transformasi untuk Prediksi

Berdasarkan analisis kebutuhan prediksi harga, data hasil transformasi akan memiliki struktur sebagai berikut:

Kolom Utama untuk Model Prediksi:

1. Target Variable (Variabel Tujuan)

- **`price_idr`** (float): Harga per malam dalam IDR (normalized)
- **`log_price`** (float): Logaritma natural dari harga (untuk transformasi)

2. Identifier Variables (Variabel Identifikasi)

- **`hotel_id`** (string): ID unik hotel hasil matching
- **`hotel_name_clean`** (string): Nama hotel yang telah distandarisasi
- **`platform_source`** (string): Sumber platform ('booking', 'tripadvisor', 'merged')

3. Geospatial Features (Fitur Geospasial)

- `latitude` (float): Koordinat latitude
- `longitude` (float): Koordinat longitude
- `distance_to_city_center` (float): Jarak ke pusat kota Yogyakarta (km)
- `distance_to_malioboro` (float): Jarak ke Jl. Malioboro (km)
- `distance_to_airport` (float): Jarak ke Bandara Yogyakarta (km)
- `distance_to_nearest_attraction` (float): Jarak ke objek wisata terdekat (km)
- `attraction_density_1km` (int): Jumlah objek wisata dalam radius 1km
- `attraction_density_2km` (int): Jumlah objek wisata dalam radius 2km
- `attraction_density_5km` (int): Jumlah objek wisata dalam radius 5km
- `restaurant_density_1km` (int): Jumlah restoran dalam radius 1km

4. Accommodation Features (Fitur Akomodasi)

- `accommodation_type` (string): Jenis akomodasi (hotel, hostel, guesthouse, etc.)
- `stars` (float): Rating bintang hotel (0-5)
- `hotel_class` (int): Kelas hotel (1-5)

5. Facilities & Amenities (Fasilitas)

- `has_wifi` (boolean): Ketersediaan WiFi gratis
- `has_breakfast` (boolean): Ketersediaan sarapan
- `has_parking` (boolean): Ketersediaan parkir
- `has_ac` (boolean): Ketersediaan AC
- `has_pool` (boolean): Ketersediaan kolam renang
- `has_gym` (boolean): Ketersediaan gym/fitness center
- `has_spa` (boolean): Ketersediaan spa
- `has_restaurant` (boolean): Ketersediaan restoran
- `has_room_service` (boolean): Ketersediaan layanan kamar
- `has_24h_reception` (boolean): Ketersediaan resepsionis 24 jam
- `facilities_score` (float): Skor agregat fasilitas (0-100)

6. Review & Rating Features (Fitur Ulasan)

- `overall_rating` (float): Rating keseluruhan (standardized 0-10)
- `rating_staff` (float): Rating staf (0-10)
- `rating_cleanliness` (float): Rating kebersihan (0-10)
- `rating_facilities` (float): Rating fasilitas (0-10)
- `rating_location` (float): Rating lokasi (0-10)
- `rating_comfort` (float): Rating kenyamanan (0-10)
- `rating_value_for_money` (float): Rating value for money (0-10)
- `total_reviews` (int): Total jumlah ulasan
- `recent_reviews` (int): Jumlah ulasan dalam 6 bulan terakhir
- `review_sentiment_score` (float): Skor sentimen agregat ulasan (-1 to 1)
- `review_consistency` (float): Konsistensi rating antar platform (0-1)

7. Competitive Features (Fitur Kompetitif)

- `competitors_within_500m` (int): Jumlah kompetitor dalam radius 500m
- `price_rank_in_area` (float): Ranking harga di area sekitar (percentile)
- `rating_rank_in_area` (float): Ranking rating di area sekitar (percentile)

8. Temporal Features (Fitur Temporal)

- `is_weekend` (boolean): Apakah akhir pekan
- `month` (int): Bulan (1-12)
- `quarter` (int): Kuartal (1-4)
- `is_peak_season` (boolean): Apakah musim puncak wisata
- `days_to_holiday` (int): Jumlah hari ke hari libur terdekat

9. Market Features (Fitur Pasar)

- `market_segment` (string): Segmen pasar (budget, mid-range, luxury)
- `target_traveler_type` (string): Tipe wisatawan target (solo, couple, family, business)
- `booking_platform_count` (int): Jumlah platform booking yang tersedia

10. Engineered Features (Fitur Hasil Rekayasa)

- `price_per_facility` (float): Harga per fasilitas yang tersedia

- `popularity_score` (float): Skor popularitas berdasarkan reviews dan rating
- `accessibility_score` (float): Skor aksesibilitas berdasarkan lokasi
- `experience_score` (float): Skor pengalaman berdasarkan ulasan
- `value_score` (float): Skor nilai berdasarkan harga vs fasilitas

11. Cross-Platform Features (Fitur Lintas Platform)

- `price_variance_platforms` (float): Varians harga antar platform
- `platform_preference_score` (float): Skor preferensi platform
- `cross_platform_availability` (boolean): Ketersediaan di multiple platform

12. Additional Metadata (Metadata Tambahan)

- `data_collection_date` (timestamp): Tanggal pengumpulan data
- `data_freshness_score` (float): Skor kesegaran data (0-1)
- `data_completeness_score` (float): Skor kelengkapan data (0-1)
- `last_updated` (timestamp): Terakhir diperbarui

Asumsi Kebutuhan Pengguna

Variabel Utama

1. Faktor Geospatial

- Jarak ke objek wisata populer (dari data geospatial)
- Jarak ke pusat kota Yogyakarta
- Kepadatan objek wisata di sekitar (radius-based analysis)

2. Fasilitas dan Layanan

- WiFi availability (dari facilities dan amenities)
- Breakfast options (dari breakfast field dan facilities)
- Room amenities (AC, TV, Kamar mandi pribadi)
- Additional services (Resepsionis 24 jam, Layanan kamar)

3. Reputasi Online

- Overall rating (dari rating field)
- Review volume (dari reviews, numberOfReviews)

- Category ratings (dari hotelRatingScores dan categoryReviewScores)
- Review sentiment (dari text analysis)

4. Cross-Platform Variables

- Price variance antar platform
- Platform availability dan preference
- Review consistency antar platform

Karakteristik Data

- **Rentang waktu**: Data terkini dengan kemampuan update berkala
- **Volume**: 100+ hotel dari kombinasi kedua platform
- **Coverage**: Akomodasi di seluruh area DIY
- **Format**: Parquet files dengan struktur tabular terstandarisasi

Tipe Data Ingestion

Tipe data ingestion yang digunakan adalah **batch processing** karena:

- Data hasil scraping dikumpulkan secara berkala
- Tidak memerlukan pemrosesan real-time
- Cocok untuk analisis pola dan tren jangka panjang
- Efisien untuk mengolah data dari berbagai sumber dengan volume besar

Tools Implementasi

Data Storage

- **Amazon S3 Bucket**: Penyimpanan raw data hasil scraping (JSON files) dan processed data (Parquet files)
- **AWS Glue Data Catalog**: Metadata management dan schema discovery

Data Processing

- **AWS Glue**: ETL jobs untuk data transformation dan integration
- **AWS Glue Crawler**: Automatic schema detection untuk raw data
- **AWS Lambda**: Trigger functions untuk batch processing

Visualization & Analysis

- **Amazon QuickSight**: Dashboard creation dan visualisasi interaktif
- **AWS Glue Studio**: Visual ETL pipeline development

Monitoring & Management

- **Amazon CloudWatch**: Monitoring pipeline performance
- **AWS CloudTrail**: Audit trail untuk data pipeline activities

Pipeline Architecture

```
[Raw Data Sources (JSON Files)]
↓
[Amazon S3 Bucket (Raw Data Storage)]
↓
[AWS Glue Crawler (Schema Discovery)]
↓
[AWS Glue Data Catalog (Metadata Management)]
↓
[AWS Glue ETL Jobs (Data Transformation)]
├─ Hotel Matching & Integration
├─ Data Cleansing & Standardization
├─ Geospatial Distance Calculation
├─ Feature Engineering
├─ Prediction Features Creation
└─ Data Quality Validation
↓
[Amazon S3 Bucket (Processed Data - Parquet Format)]
├─ hotel_features.parquet
├─ review_features.parquet
├─ geospatial_features.parquet
└─ prediction_dataset.parquet
↓
[Amazon QuickSight (Visualization & Dashboard)]
```

Struktur File Output Parquet

1. prediction_dataset.parquet

File utama untuk model prediksi yang berisi semua kolom yang diperlukan untuk training dan inference.

2. hotel_features.parquet

Data fitur hotel termasuk fasilitas, rating, dan karakteristik dasar.

3. review_features.parquet

Data fitur ulasan termasuk sentimen, rating kategori, dan statistik ulasan.

4. geospatial_features.parquet

Data fitur geospasial termasuk jarak ke berbagai landmark dan density analysis.

Manfaat Pipeline Data

Pipeline data ini bermanfaat bagi:

- **Pemilik akomodasi**: Strategi penetapan harga kompetitif dan optimasi fasilitas
- **Investor pariwisata**: Analisis pasar dan identifikasi peluang investasi
- **Wisatawan**: Informasi perbandingan harga dan fasilitas optimal
- **Pengelola pariwisata daerah**: Pemahaman dinamika pasar akomodasi DIY
- **Pemerintah daerah**: Data pendukung pengembangan kebijakan pariwisata
- **Platform booking**: Market intelligence dan competitive analysis
- **Data Scientists**: Dataset yang siap digunakan untuk machine learning dan analisis prediktif

Catatan Implementasi

Semua tools yang dipilih tersedia dalam AWS Free Tier atau AWS Student Account, memastikan biaya implementasi yang minimal sambil tetap memberikan fungsionalitas penuh untuk pengembangan pipeline data yang robust dan scalable. Output dalam format Parquet di S3 memudahkan integrasi dengan berbagai tools analisis dan machine learning frameworks.

Validasi Data dan Quality Assurance

Data Quality Metrics

- **Completeness**: Persentase data yang tidak kosong per kolom
- **Accuracy**: Validasi format dan range nilai
- **Consistency**: Konsistensi data antar platform
- **Timeliness**: Kesegaran data dan frekuensi update
- **Uniqueness**: Deteksi duplikasi data

Automated Testing

- Unit testing untuk setiap fungsi transformasi
- Integration testing untuk pipeline end-to-end

- Data validation rules untuk setiap tahap transformasi
- Alerting system untuk anomali data atau pipeline failure

Dokumen ini memberikan panduan lengkap untuk implementasi pipeline data yang dapat menghasilkan dataset berkualitas tinggi untuk prediksi harga akomodasi di Yogyakarta.