**LAPORAN TUGAS BESAR BIG DATA**

****

**Project Based Assessment**

**Dataset 2015 Flight Delay**

Disusun oleh:

1. Andi Muhammad Naufal 1301223340
2. Jevon Sebastian 1301223391

**PROGRAM STUDI S1 INFORMATIKA**

**FAKULTAS INFORMATIKA**

**TELKOM UNIVERSITY**

**2025**

### 

### **Daftar Isi**

1. Pendahuluan

1.1. Latar Belakang

1.2. Rumusan Masalah

1.3. Tujuan Analisis

1.4. Ruang Lingkup

2. Kajian Pustaka

2.1. Analisis Deret Waktu (*Time Series*)

2.2. Model SARIMA (Seasonal Autoregressive Integrated Moving Average)

2.3. *Big Data* dan Pemrosesan Terdistribusi

2.4. Apache Spark (PySpark)

3. Metodologi Analisis

3.1. Alur Kerja Proyek

3.2. Deskripsi Data

3.3. Lingkungan Teknis dan Pustaka

3.4. Pra-pemrosesan dan Transformasi Data dengan PySpark

3.5. Analisis Eksploratif dan Dekomposisi Deret Waktu

3.6. Pemodelan dengan SARIMA

3.7. Metrik Evaluasi Model

4. Hasil dan Pembahasan

4.1. Hasil Pra-pemrosesan Data

4.2. Visualisasi dan Dekomposisi Volume Penerbangan Harian

4.3. Pembagian Data Latih dan Uji

4.4. Pemilihan dan Pelatihan Model SARIMA

4.5. Evaluasi Kinerja Model

4.6. Proyeksi (Forecasting) Volume Penerbangan

5. Kesimpulan dan Rekomendasi

5.1. Kesimpulan

5.2. Keterbatasan Analisis

5.3. Rekomendasi

### 1. Pendahuluan

1.1. Latar Belakang

Industri penerbangan merupakan salah satu sektor paling dinamis dan kompetitif di dunia. Volume lalu lintas udara tidak hanya menjadi indikator penting bagi kesehatan ekonomi global, tetapi juga menjadi dasar bagi perencanaan operasional dan strategis maskapai penerbangan, pengelola bandara, dan bisnis terkait lainnya. Fluktuasi jumlah penerbangan harian dipengaruhi oleh berbagai faktor, termasuk hari dalam seminggu, musim liburan, kondisi ekonomi, hingga peristiwa tak terduga.

Memahami pola historis dan mampu memprediksi volume penerbangan di masa depan dengan akurat adalah sebuah keunggulan kompetitif. Prediksi yang andal memungkinkan optimalisasi alokasi sumber daya, seperti penjadwalan kru, manajemen armada pesawat, perencanaan kapasitas bandara, dan strategi penetapan harga tiket. Dengan data historis penerbangan yang tersedia dalam volume sangat besar (Big Data), pemanfaatan teknologi pemrosesan data canggih dan metode peramalan statistik yang kuat menjadi suatu keharusan.

1.2. Rumusan Masalah

Analisis ini berupaya menjawab pertanyaan-pertanyaan berikut:

1. Bagaimana pola historis volume penerbangan harian di Amerika Serikat selama tahun 2015?
2. Bagaimana cara mengelola dan memproses dataset penerbangan berskala besar secara efisien untuk analisis deret waktu?
3. Model peramalan deret waktu apakah yang cocok untuk memprediksi volume penerbangan harian dengan mempertimbangkan pola musiman (mingguan)?
4. Seberapa akurat model yang dibangun dalam memprediksi volume penerbangan pada periode yang tidak terlihat sebelumnya?

1.3. Tujuan Analisis

Tujuan utama dari laporan ini adalah:

1. Melakukan pra-pemrosesan dataset penerbangan mentah yang berukuran besar menggunakan teknologi *Big Data* (Apache Spark) untuk menghasilkan data deret waktu yang siap dianalisis.
2. Mengidentifikasi tren dan pola musiman dari data volume penerbangan harian.
3. Membangun, melatih, dan mengevaluasi model peramalan *Seasonal Autoregressive Integrated Moving Average* (SARIMA) untuk memprediksi jumlah penerbangan harian.
4. Menghasilkan proyeksi volume penerbangan untuk 30 hari ke depan sebagai dasar pengambilan keputusan.

1.4. Ruang Lingkup

Analisis ini menggunakan dataset publik yang mencakup detail penerbangan domestik di Amerika Serikat untuk tahun 2015. Fokus analisis adalah pada agregasi jumlah total penerbangan per hari. Proses analisis mencakup pembersihan data, transformasi menggunakan PySpark, analisis eksploratif, pemodelan menggunakan SARIMA, dan evaluasi performa model. Proyeksi masa depan dibatasi hingga 30 hari setelah periode data yang tersedia.

### 2. Kajian Pustaka

2.1. Analisis Deret Waktu (*Time Series*)

Analisis deret waktu adalah serangkaian teknik statistik yang digunakan untuk menganalisis data yang diindeks dalam urutan waktu. Tujuan utamanya adalah untuk mengidentifikasi pola—seperti tren, musiman, dan siklus—untuk memahami perilaku data di masa lalu dan membuat prediksi (forecast) tentang masa depan.

2.2. Model SARIMA (Seasonal Autoregressive Integrated Moving Average)

SARIMA adalah salah satu model peramalan deret waktu yang paling populer dan kuat, terutama untuk data yang menunjukkan pola musiman. Model ini merupakan pengembangan dari model ARIMA dengan menambahkan komponen musiman.

Model SARIMA dijelaskan dengan notasi SARIMA(p,d,q)(P,D,Q)\_m, di mana:

* (p, d, q) adalah parameter non-musiman:
  + p (Autoregressive): Jumlah observasi lampau yang dimasukkan ke dalam model.
  + d (Integrated): Jumlah proses *differencing* yang diperlukan untuk membuat data stasioner.
  + q (Moving Average): Orde dari rata-rata bergerak, merepresentasikan jumlah kesalahan prediksi lampau dalam model.
* (P, D, Q, m) adalah parameter musiman:
  + P, D, Q: Entitas yang sama dengan (p, d, q) tetapi untuk komponen musiman.
  + m: Jumlah periode dalam satu musim (misalnya, m=7 untuk data harian dengan siklus mingguan).

2.3. *Big Data* dan Pemrosesan Terdistribusi

*Big Data* merujuk pada kumpulan data yang sangat besar dan kompleks sehingga sulit untuk dikelola dan diproses menggunakan aplikasi pemrosesan data tradisional. Dataset penerbangan, yang dapat berisi jutaan catatan, termasuk dalam kategori ini. Pemrosesan terdistribusi adalah pendekatan di mana tugas komputasi dibagi menjadi bagian-bagian kecil dan didistribusikan ke beberapa komputer (atau *core* prosesor) dalam sebuah klaster untuk dieksekusi secara paralel, sehingga meningkatkan kecepatan dan efisiensi secara drastis.

2.4. Apache Spark (PySpark)

Apache Spark adalah kerangka kerja komputasi terdistribusi sumber terbuka yang dirancang untuk pemrosesan data berskala besar. Spark menyediakan antarmuka pemrograman untuk seluruh klaster dengan toleransi kesalahan (*fault tolerance*) implisit. PySpark adalah API Python untuk Spark, yang memungkinkan para ilmuwan data dan insinyur untuk memanfaatkan kekuatan Spark dalam lingkungan Python yang familiar. Dalam analisis ini, PySpark digunakan untuk tahap pra-pemrosesan data mentah, di mana efisiensi dalam menangani jutaan baris data sangat krusial.

### 3. Metodologi Analisis

3.1. Alur Kerja Proyek

Proyek ini mengikuti alur kerja standar dalam ilmu data, yang diadaptasi untuk menangani dataset besar dan analisis deret waktu.

1. Inisialisasi Lingkungan: Pengaturan lingkungan komputasi dan instalasi pustaka yang diperlukan seperti PySpark, Pandas, Statsmodels, dan Pmdarima.
2. Akuisisi dan Pemuatan Data: Data mentah flights.csv dimuat ke dalam DataFrame Apache Spark untuk pemrosesan awal.
3. Pra-pemrosesan dengan PySpark:
   * Penyaringan data untuk mengecualikan penerbangan yang dibatalkan (*cancelled*) atau dialihkan (*diverted*).
   * Pembuatan kolom tanggal tunggal (flight\_date) dari kolom YEAR, MONTH, dan DAY.
   * Agregasi data untuk menghitung total penerbangan per hari (num\_flights).
   * Pengurutan data berdasarkan tanggal.
4. Konversi ke Pandas: Hasil agregasi yang sudah jauh lebih kecil dikonversi menjadi Pandas DataFrame untuk analisis deret waktu lebih lanjut.
5. Analisis dan Visualisasi: Visualisasi data deret waktu untuk mengidentifikasi pola dan melakukan dekomposisi untuk memisahkan komponen tren, musiman, dan residual.
6. Pembagian Data: Data dibagi menjadi set pelatihan (Januari - Oktober 2015) dan set pengujian (November - Desember 2015).
7. Pemodelan SARIMA:
   * Menggunakan fungsi auto\_arima untuk secara otomatis menemukan parameter SARIMA (p,d,q)(P,D,Q)\_m terbaik.
   * Melatih model SARIMA pada set data pelatihan menggunakan parameter yang ditemukan.
8. Evaluasi Model: Model yang telah dilatih digunakan untuk membuat prediksi pada set data pengujian. Kinerja prediksi dievaluasi menggunakan metrik MAE, RMSE, dan MAPE.
9. Forecasting: Model dilatih ulang menggunakan keseluruhan data untuk membuat prediksi 30 hari ke depan.

3.2. Deskripsi Data

Dataset yang digunakan adalah flights.csv, yang berisi informasi detail penerbangan di AS. Dataset ini memiliki 5.819.079 baris dan 31 kolom. Kolom-kolom kunci yang digunakan dalam analisis ini meliputi:

* YEAR, MONTH, DAY: Komponen tanggal penerbangan.
* CANCELLED: Indikator apakah penerbangan dibatalkan.
* DIVERTED: Indikator apakah penerbangan dialihkan.

3.3. Lingkungan Teknis dan Pustaka

Analisis dilakukan dalam lingkungan berbasis Python menggunakan Google Colab. Pustaka utama yang digunakan adalah:

* PySpark: Untuk pemrosesan data skala besar.
* Pandas: Untuk manipulasi data deret waktu setelah agregasi.
* NumPy: Untuk operasi numerik.
* Matplotlib & Seaborn: Untuk visualisasi data.
* Statsmodels: Untuk dekomposisi deret waktu dan pemodelan SARIMA.
* Pmdarima: Untuk pencarian otomatis parameter SARIMA terbaik.

3.4. Pra-pemrosesan dan Transformasi Data dengan PySpark

Mengingat volume data yang besar, Pandas read\_csv akan memakan banyak memori dan waktu. Oleh karena itu, langkah awal dilakukan menggunakan PySpark spark.read.csv yang lebih efisien. Transformasi data dilakukan secara terdistribusi:

1. Penerbangan yang valid (tidak dibatalkan atau dialihkan) dipilih menggunakan df\_spark.filter().
2. Kolom tanggal (flight\_date) dibuat dengan menggabungkan kolom YEAR, MONTH, dan DAY.
3. Agregasi data dilakukan dengan groupBy('flight\_date').count() untuk menghitung jumlah penerbangan harian.

Langkah ini secara efektif mereduksi jutaan catatan penerbangan menjadi satu data deret waktu harian, yang kemudian dapat dianalisis dengan nyaman menggunakan Pandas.

3.5. Analisis Eksploratif dan Dekomposisi Deret Waktu

Setelah data agregat tersedia, dilakukan visualisasi untuk mengamati tren dan musiman. Dekomposisi aditif digunakan untuk memecah deret waktu menjadi tiga komponen:

* Tren (Trend): Pergerakan data jangka panjang.
* Musiman (Seasonal): Pola berulang yang sistematis (dalam kasus ini, mingguan).
* Sisaan (Residual): Komponen acak setelah tren dan musiman dihilangkan.

3.6. Pemodelan dengan SARIMA

Proses pemodelan melibatkan dua tahap utama. Pertama, fungsi auto\_arima dari pustaka pmdarima digunakan pada data pelatihan. Fungsi ini secara cerdas menguji berbagai kombinasi parameter SARIMA dan memilih yang terbaik berdasarkan kriteria informasi seperti AIC (*Akaike Information Criterion*). Musiman mingguan secara eksplisit didefinisikan dengan parameter m=7.

Setelah parameter optimal ditemukan, model SARIMAX dari statsmodels dilatih menggunakan parameter tersebut pada data pelatihan.

3.7. Metrik Evaluasi Model

Untuk mengukur seberapa baik model memprediksi data yang belum pernah dilihatnya (data uji), tiga metrik standar digunakan:

1. Mean Absolute Error (MAE): Rata-rata dari nilai absolut kesalahan prediksi. Memberikan gambaran besaran kesalahan dalam unit yang sama dengan data asli.
2. Root Mean Squared Error (RMSE): Akar kuadrat dari rata-rata kuadrat kesalahan. Metrik ini memberikan bobot lebih besar pada kesalahan yang besar.
3. Mean Absolute Percentage Error (MAPE): Rata-rata persentase kesalahan absolut. Metrik ini berguna untuk memahami skala kesalahan relatif terhadap nilai aktual.

### 4. Hasil dan Pembahasan

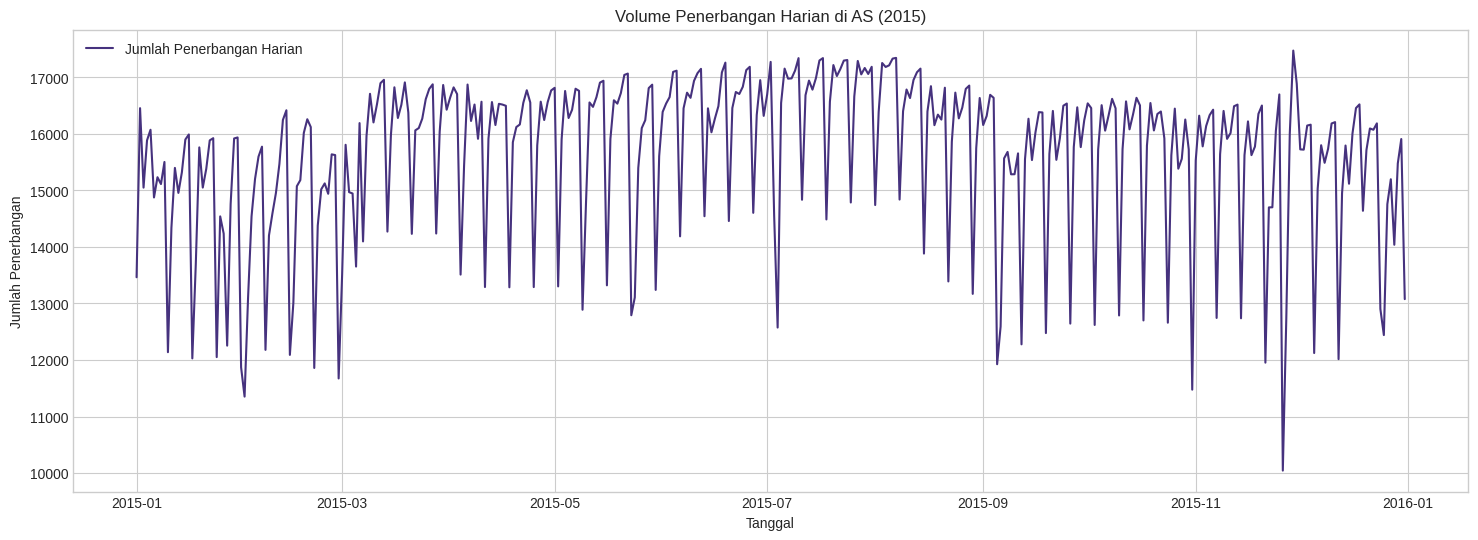
4.1. Hasil Pra-pemrosesan Data

Penggunaan PySpark berhasil mengelola dataset mentah berukuran lebih dari 1.3 GB. Proses pemfilteran, pembuatan fitur tanggal, dan agregasi berjalan secara efisien. Hasil akhir dari blok pemrosesan ini adalah sebuah DataFrame Pandas (daily\_flights\_volume\_ts) yang berisi dua kolom: flight\_date (sebagai indeks) dan num\_flights (jumlah penerbangan). Data agregat pertama menunjukkan 13.464 penerbangan pada 1 Januari 2015.

4.2. Visualisasi dan Dekomposisi Volume Penerbangan Harian

Grafik volume penerbangan harian sepanjang tahun 2015 (Gambar 1) menunjukkan pola yang sangat jelas dan berulang.

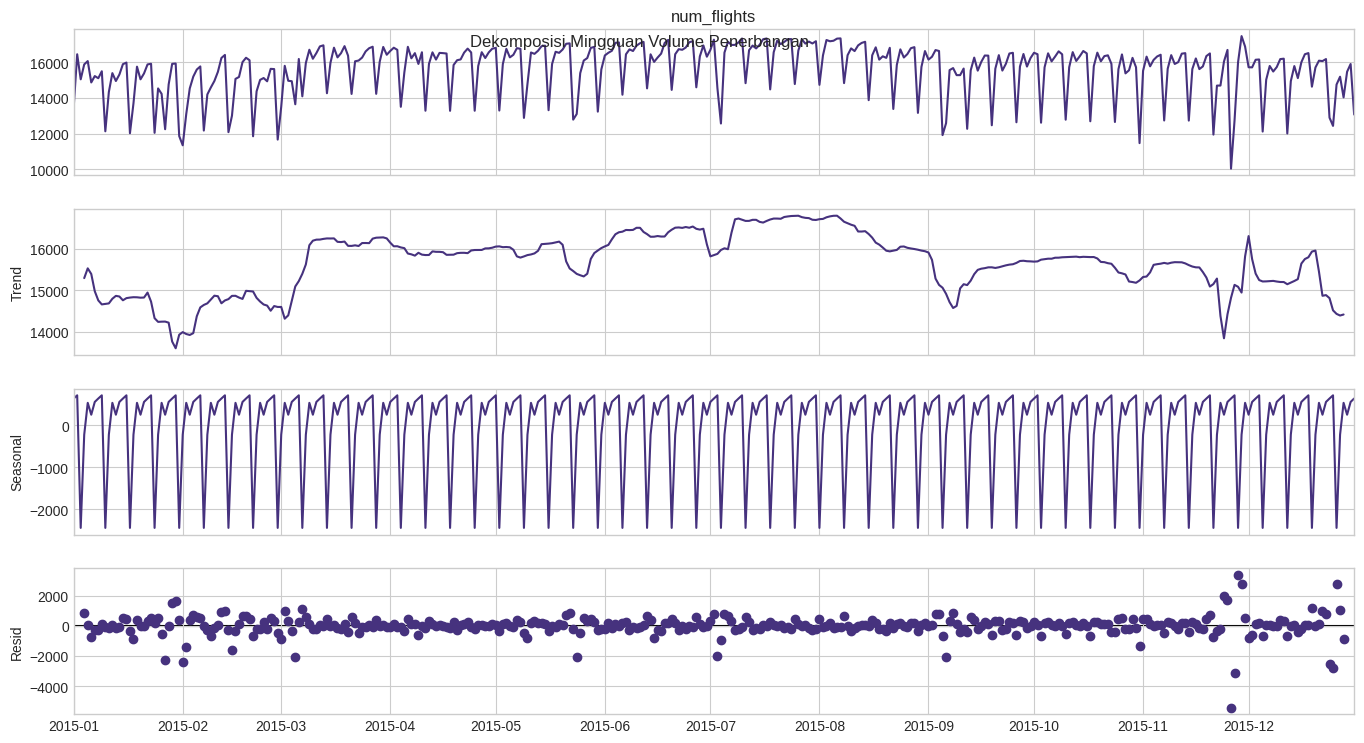
*Gambar 1: Volume Penerbangan Harian di AS (2015)*

**

Dekomposisi deret waktu (Gambar 2) mengonfirmasi pengamatan ini dan memecahnya menjadi komponen-komponen yang lebih jelas:

* Observed: Data asli.
* Trend: Komponen tren menunjukkan adanya sedikit peningkatan volume penerbangan dari awal tahun hingga pertengahan tahun (sekitar bulan Agustus), kemudian sedikit menurun menjelang akhir tahun. Ini mungkin berkorelasi dengan musim liburan musim panas.
* Seasonal: Komponen musiman menunjukkan pola mingguan yang sangat kuat dan konsisten. Volume penerbangan mencapai puncaknya pada hari-hari kerja dan turun secara signifikan selama akhir pekan.
* Residual: Komponen sisaan tampak relatif acak, menunjukkan bahwa model tren dan musiman telah berhasil menangkap sebagian besar pola dalam data.

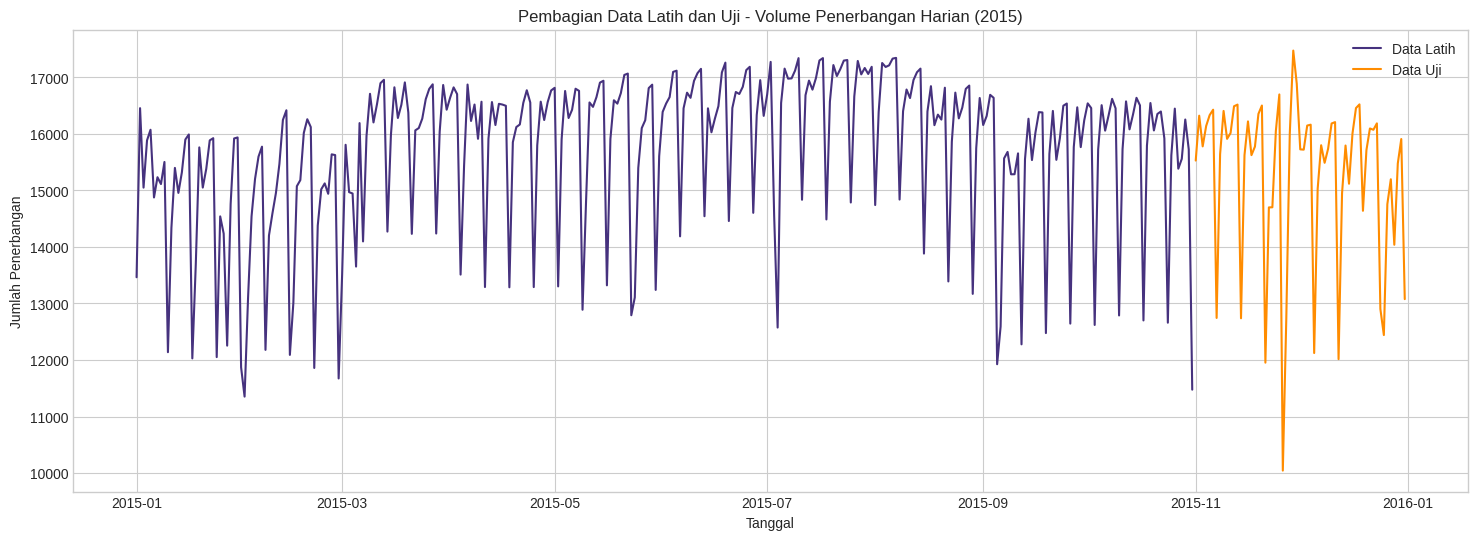
*Gambar 2: Dekomposisi Mingguan Volume Penerbangan*

**

4.3. Pembagian Data Latih dan Uji

Data dibagi untuk tujuan pelatihan dan validasi. Sebanyak 304 observasi (data dari 1 Januari hingga 31 Oktober 2015) digunakan sebagai data latih, dan 61 observasi sisanya (1 November hingga 31 Desember 2015) digunakan sebagai data uji. Pembagian ini memastikan bahwa model dievaluasi kemampuannya untuk melakukan peramalan "out-of-sample".

*Gambar 3: Pembagian Data Latih dan Uji*

**

4.4. Pemilihan dan Pelatihan Model SARIMA

Proses auto\_arima dijalankan pada data pelatihan untuk mencari parameter terbaik. Hasilnya adalah sebagai berikut:

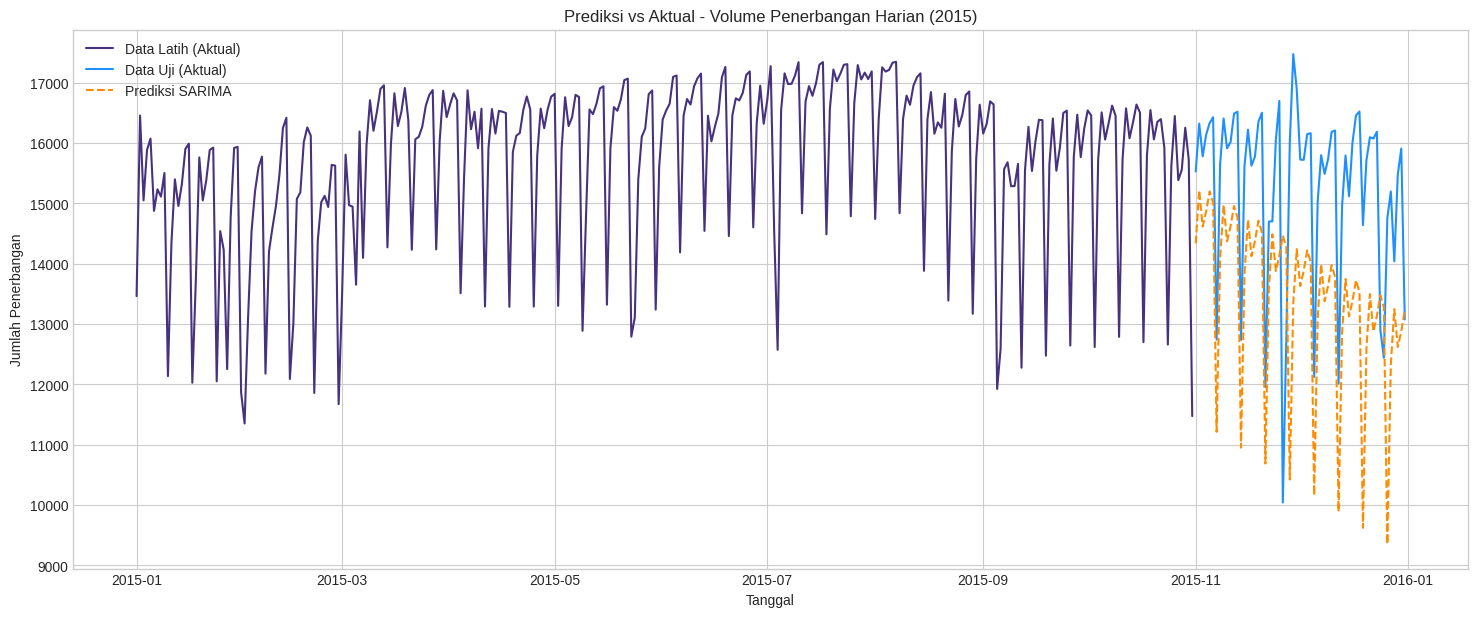
* Parameter non-musiman (p,d,q): (0, 1, 0)
* Parameter musiman (P,D,Q,m): (1, 0, 1, 7)

Model SARIMA(0,1,0)(1,0,1)₇ kemudian dilatih pada data pelatihan. Plot diagnostik yang dihasilkan (tidak ditampilkan secara detail di sini, namun disebutkan dalam kode) digunakan untuk memvalidasi asumsi model pada data latih, seperti normalitas dan tidak adanya korelasi pada residual.

4.5. Evaluasi Kinerja Model

Model yang telah dilatih digunakan untuk memprediksi volume penerbangan pada data uji (November-Desember 2015). Gambar 4 membandingkan nilai prediksi (garis putus-putus oranye) dengan nilai aktual pada data latih dan uji.

*Gambar 4: Prediksi SARIMA vs Nilai Aktual*

**

Secara visual, prediksi model berhasil menangkap pola musiman mingguan dengan baik, meskipun terdapat beberapa deviasi pada puncak dan lembah absolut. Kinerja kuantitatif model pada data uji adalah sebagai berikut:

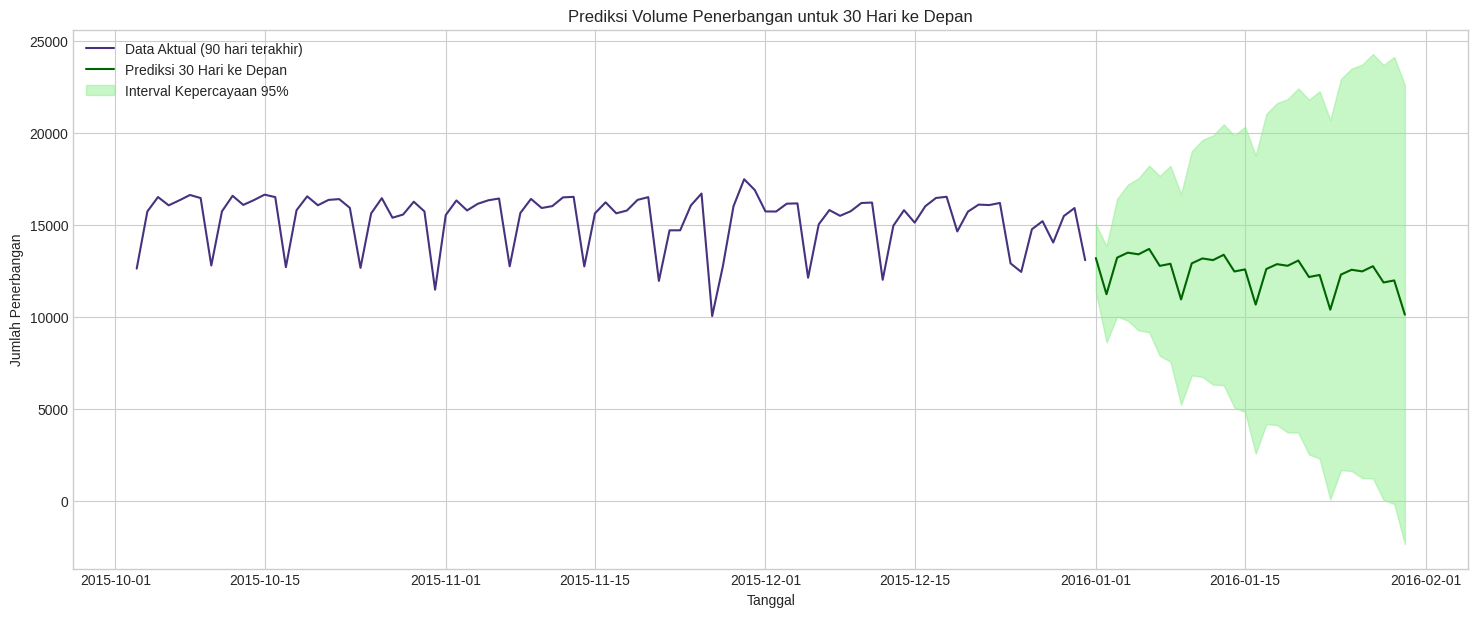
* Mean Absolute Error (MAE): 2096.81 penerbangan
* Root Mean Squared Error (RMSE): 2363.69 penerbangan
* Mean Absolute Percentage Error (MAPE): 13.83%

MAE sebesar ~2097 penerbangan menunjukkan bahwa, secara rata-rata, prediksi model meleset sekitar 2097 penerbangan dari jumlah aktual harian. MAPE sebesar 13.83% memberikan konteks persentase dari kesalahan tersebut. Mengingat variabilitas harian yang tinggi (rentang antara ~10.000 hingga ~17.000 penerbangan), kinerja ini dapat dianggap cukup baik sebagai model dasar.

4.6. Proyeksi (Forecasting) Volume Penerbangan

Untuk membuat proyeksi yang paling andal, model SARIMA dilatih ulang menggunakan seluruh data yang tersedia (Januari hingga Desember 2015). Kemudian, model ini digunakan untuk meramalkan volume penerbangan selama 30 hari ke depan.

*Gambar 5: Proyeksi Volume Penerbangan untuk 30 Hari ke Depan*

**

Hasil proyeksi (garis hijau tua) dengan jelas melanjutkan pola musiman mingguan yang diamati pada data historis. Area hijau muda menunjukkan interval kepercayaan 95%, yang merepresentasikan rentang di mana nilai aktual kemungkinan besar akan jatuh. Interval ini melebar seiring berjalannya waktu, yang mencerminkan meningkatnya ketidakpastian pada prediksi yang lebih jauh ke masa depan. Proyeksi ini dapat memberikan wawasan berharga bagi pemangku kepentingan untuk perencanaan jangka pendek.

### 5. Kesimpulan dan Rekomendasi

5.1. Kesimpulan

Analisis ini telah berhasil menunjukkan alur kerja lengkap untuk memprediksi volume penerbangan harian dari dataset mentah berskala besar.

1. Penggunaan PySpark terbukti sangat efektif untuk tahap pra-pemrosesan data, memungkinkan agregasi efisien dari jutaan catatan penerbangan menjadi deret waktu harian yang dapat dikelola.
2. Volume penerbangan harian di AS pada tahun 2015 menunjukkan pola musiman mingguan yang sangat kuat, dengan lalu lintas yang lebih rendah pada akhir pekan, serta tren tahunan yang moderat.
3. Model SARIMA(0,1,0)(1,0,1)₇ berhasil diidentifikasi dan dilatih untuk menangkap struktur data ini.
4. Model menunjukkan kinerja prediksi yang layak pada data uji, dengan MAPE sebesar 13.83%, dan mampu menghasilkan proyeksi 30 hari ke depan yang secara logis melanjutkan pola historis.

5.2. Keterbatasan Analisis

Analisis ini memiliki beberapa keterbatasan yang perlu diperhatikan:

* Data Satu Tahun: Model dibangun hanya berdasarkan data dari satu tahun (2015). Ini membatasi kemampuan model untuk menangkap tren jangka panjang atau musiman tahunan yang lebih kompleks (misalnya, pengaruh hari libur besar seperti Thanksgiving atau Natal yang jatuh pada hari yang berbeda setiap tahunnya).
* Variabel Eksogen: Model ini bersifat univariat, artinya hanya menggunakan data historis volume penerbangan itu sendiri. Model tidak memasukkan variabel eksternal (eksogen) yang dapat mempengaruhi volume penerbangan, seperti harga bahan bakar, kondisi cuaca, hari libur nasional, atau indikator ekonomi.

5.3. Rekomendasi

Berdasarkan temuan dan keterbatasan, berikut adalah beberapa rekomendasi untuk pengembangan di masa depan:

1. Perluasan Dataset: Menggunakan data dari beberapa tahun akan memungkinkan model untuk mempelajari pola tahunan dan tren jangka panjang dengan lebih baik, sehingga meningkatkan akurasi.
2. Model SARIMAX: Mengembangkan model menjadi SARIMAX (SARIMA with eXogenous variables). Memasukkan fitur-fitur seperti indikator hari libur, hari dalam seminggu, dan mungkin data cuaca dapat secara signifikan meningkatkan akurasi prediksi.
3. Eksplorasi Model Alternatif: Mengeksplorasi model-model peramalan lain seperti Prophet (dikembangkan oleh Facebook) atau model berbasis *machine learning* (misalnya, LSTM atau Gradient Boosting) dapat memberikan perbandingan kinerja dan mungkin menghasilkan hasil yang lebih baik.
4. Analisis Per Rute atau Bandara: Melakukan analisis serupa pada tingkat yang lebih granular, seperti per rute penerbangan atau per bandara, dapat memberikan wawasan operasional yang lebih spesifik dan dapat ditindaklanjuti bagi maskapai dan pengelola bandara.