Alur Pre-processing Data AIS

Langkah	Nama Tahap	Tujuan Singkat	Kriteria Lulus
1	**Validasi Skema & Kelengkapan**	Pastikan kolom inti tersedia (`mmsi`, `dt_pos_utc`, `lat`, `lon`, `sog`) & bertipe benar	Kolom hilang ⇢ file ditolak tipe salah ⇢ cas jika gagal ⇢ drop baris
			Valid MMSI = MMSI < 100000000 or MMSI > 999999999 Valid IMO = 1000000 - 9999999 Valid lat = rentang **-90° hingga +90°** adalah rentang penuh dari kutub selatan ke kutub utara Valid lot = rentang **-180° sampai 180°** dan mengukur posisi timur-barat terhadap **meridian utama (Greenwich)**
2	**Deduplikasi Prime Key**	Hilangkan rekaman ganda	Definisi PK = `(mmsi, dt_pos_utc[, message_type])`; duplikat ⇢ simpan satu, log sisanya
3	**Filter Domain (Rules Hard)**	Bersihkan *outlier* eksplisit	• MMSI 9-digit valid • `draught ≠ 0` • Tipe kapal

			relevan (cargo, tanker) • Periode analisis (mis. 2022)
4	**Pengujian Range Numerik**	Jaga batas fisik & format	`lat` ±90°, `lon` ±180°; `0 ≤ sog ≤ 70 kn`; timestamp ≤ _UTC now + 5 m
5	**Konsistensi Metadata per MMSI**	Satu identitas kapal = satu set metadata	Hitung `nunique()` `imo`, `vessel_name`, `length`, `width`; >1 ⇢ flag audit
6	**Koherensi Spasio-Temporal**	Deteksi loncatan & jeda sinyal	• *Jump test* > 30 km/min ⇢ drop titik • `Δt > 60 m` ⇢ flag gap • `nav_status` ↔ `sog` selaras
7	**Enrichment & Spatial Join**	Tambahkan konteks	• Join data statis kapal (IMO, tipe) • Spatial join dengan poligon pelabuhan → `inside_port`

QA Quality Dimension	Indicator QA																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Relevance			v			v			v								v	
Accuracy	v	v	v		v		v	v		v	v	v	v	v	v	v		v
Punctuality and Timeliness												v						
Interpretability							v											
Accessibility																		
Coherence		v		v	v	v		v						v		v		v
Comparability																		
Trustworthiness	v									v								

Kerangka Data-Input Quality Metrics (AIS):

	Indikator	Dimensi Kualitas	Goals (penjelasan singkat tapi jelas)	Rumus Indikator*	Urgensi
1	Percent of Missing Values	Accuracy, Trustworthiness	Memastikan kolom-kolom kunci (`mmsi`, waktu, koordinat, kecepatan) selalu terisi, supaya lintasan kapal tidak “bolong-bolong” dan analisis port-call tidak salah sasaran.	`null_pct(col) = (Σ row_null(col) / N_total) × 100 %`	Critical
2	Invalid MMSI Rate	Accuracy, Coherence	Menyingkirkan pesan AIS yang memakai MMSI di luar 9 digit resmi—karena rekaman palsu dapat “mencemari” statistik per kapal dan pelabuhan.	invalid_mmsi_pct = (Σ MMSI < 100 000 000 ∨ > 999 999 999) / N_total × 100 %	Critical

3	Non-Zero Draught Rate	Relevance, Accuracy	Sarat air 0 artinya nakhoda tidak mengisi / sensor mati. Kita butuh nilai >0 untuk menganalisis bobot muatan dan kedalaman pelabuhan.	`draught_zero_pct = (Σ draught = 0) / N_total × 100 %`	Important
4	Records & Unique Vessels	Coverage, Coherence	Mengecek apakah jumlah total ping dan kapal harian/-mingguan berada di kisaran normal; penurunan tajam bisa menandakan gangguan satelit atau ingest.	`rows_cnt`, `mmsi_n = COUNT(DISTINCT mmsi)` dibandingkan baseline rata-rata	Critical
5	Avg. Records per Vessel	Coherence, Accuracy	Mengukur kepadatan ping tiap kapal; kalau terlalu sedikit, kita bisa kehilangan detil manuver di pelabuhan.	`avg_rows_per_mmsi = rows_cnt / mmsi_n`	Important
6	Day-of-Week Pattern	Coherence, Relevance	Melihat distribusi lalu-lintas Senin → Minggu; pola “kosong di Minggu” misalnya bisa ikut memengaruhi jam kerja pelabuhan.	Hitung `rows_cnt` & `mmsi_n` per hari (Mon…Sun); cek hari dengan 0 rekor	Nice

7	Schema & Range Pass-Rate	Accuracy, Interpretability	Memastikan semua kolom berada di rentang fisik (lat ±90°, lon ±180°, sog 0-70 kn, dst); baris di luar itu dianggap cacat.	`pass_rate = rows_good / N_total × 100 %`	Critical
8	Unique H3 Cells (res 7)	Coverage, Accuracy	Menguji apakah cakupan spasial data sudah menutupi area laut yang semestinya; sel res 7 ≈ 1 km².	`h3_unique_cnt = COUNT(DISTINCT h3_r7)`	Important
9	Ports Covered per Day	Coverage, Relevance	Menjamin data mencakup cukup banyak pelabuhan; penurunan drastis bisa berarti layer poligon salah atau data hilang.	`port_n_daily = COUNT(DISTINCT port_id)`	Important
10	Incorrect Coordinates	Accuracy, Trustworthiness	Menolak titik AIS yang “mendarat” di daratan atau di luar peta laut; ini biasanya noise GPS atau spoofing.	`coord_on_land_pct = (Σ point_on_land) / N_total × 100 %`	Critical

Commented [1]: Penghitungannya mungkin dibagi per orang per port nanti kalau pakai ini takutnya gak dapet ? gimana han ?

11	Spatial Gap (Jump Test)	Accuracy	Mengidentifikasi “teleport” > 30 km per menit yang mustahil secara fisik sehingga wajib dibuang.	`jump_rate = distance_km/Δt_min`; flag jika > 30	Critical
12	Time Gap between Pings	Timeliness, Accuracy	Mencari jeda sinyal > 60 menit; kalau banyak, artinya ada blank spot satelit atau AIS dimatikan.	Distribusi `Δt = t_i − t_{i-1}`; hitung % gap > 60 m	Important
13	Speed Gap Consistency	Accuracy	Memastikan kecepatan terhitung (jarak/Δt) sejalan dengan `sog`; gap besar ≈ sensor salah.	`speed_calc_kmh = (distance_km / Δt_h)`, lalu	calc - sog
14	Metadata Consistency	Coherence, Accuracy	`imo`, nama, ukuran kapal seharusnya tidak berubah-ubah dalam satu MMSI; perubahan = sinyal kesalahan merge.	`meta_conflict = (nunique(col) > 1)` per mmsi	Important
15	Ingest Latency p95	Timeliness	Menilai apakah data masuk hampir real-time; latency tinggi mengurangi	`latency_p95 = p95(dt_insert - dt_pos)`	Nice

			manfaat operasional.		
16	Duplicate Primary-Key Rate	Accuracy, Coherence	(mmsi, timestamp) wajib unik; duplikat artinya data direplay atau file dobel.	`dup_pk_pct = (Σ dup_pk) / N_total × 100 %`	Critical
17	Inside-Port Share	Relevance	Memilih kapal yang betul-betul pernah berada di poligon pelabuhan—hanya mereka yang relevan untuk statistik port-call.	`inside_port_share_mmsi = rows_inside / rows_total`	Important
18	Nav-Status vs Speed	Coherence, Accuracy	Mengecek kecocokan status ‘Moored/Anchored’ dengan kecepatan < 0,5 kn; inkonsistensi = potensi kesalahan sensor.	`incoherent_nav_pct = (Σ rows_nav_mismatch)/ N_total × 100 %`	Nice