HPCL B2B Lead Intelligence System

Product Intelligence Dataset - Technical Documentation

Version 1.0 | February 2026

Table of Contents

1. Executive Summary

1.1 Purpose

This dataset powers HPCL's B2B Lead Intelligence System, enabling automated detection and qualification of potential customers from public signals. It translates unstructured information from the internet into actionable, scored leads for the Direct Sales team.

1.2 Scope

The dataset covers **12 product categories** across HPCL's Direct Sales portfolio:

- High Speed Diesel (HSD) - Captive power & DG sets
- Light Diesel Oil (LDO) - Small gensets & burners
- Furnace Oil (FO) - Industrial boilers & furnaces
- Bitumen - Road construction & infrastructure
- Marine Bunker Fuels - Shipping & port operations
- Hexane - Solvent extraction & chemical processing
- Propylene - Petrochemical feedstock
- Jute Batch Oil (JBO) - Jute processing
- Solvent 1425 - Paint & coating manufacturing
- Mineral Turpentine Oil (MTO 2445) - Industrial solvents
- Sulphur - Fertilizer & chemical production
- Superior Kerosene Oil Non-PDS - Industrial applications

1.3 Key Capabilities

- **Multi-layered Signal Detection:** Primary keywords, secondary keywords, contextual phrases, negative filters, and tender-specific patterns
- **Industry Intelligence:** High/medium confidence industry mappings with sector-specific signals
- **Dynamic Confidence Scoring:** 0.30-0.95 scale with multi-factor scoring and disqualifiers
- **Cross-Product Intelligence:** Automatic detection of multi-product opportunities
- **Temporal Awareness:** Seasonal patterns, budget cycles, and event-driven demand
- **Quality Gates:** Automated validation, deduplication, and human review thresholds

2. Dataset Architecture

2.1 Top-Level Structure

The dataset is a JSON document with the following root-level objects:

Object	Purpose
metadata	Version control, ownership, and update tracking
products	Array of 12 product configurations (core dataset)
crossProductIntelligence	Rules for multi-product opportunity detection
verificationLayer	Quality control thresholds and validation rules
alertingRules	Lead routing and notification logic
feedbackLoop	Continuous learning and model improvement
sourceRegistry	Permitted data sources and access methods
complianceAndGovernance	Data privacy and ethical guidelines

2.2 Data Flow

The system processes signals through this pipeline:

10. Web Intelligence Layer → Signal extracted from permitted sources
11. Entity Resolution → Company identity normalized and unified
12. Signal Detection → Keywords and patterns matched against product configs
13. Industry Mapping → Company classified into high/medium confidence industries
14. Inference Engine → Product needs inferred with reason codes
15. Confidence Scoring → Multi-factor score calculated (0.30-0.95)
16. Validation Layer → Quality checks and deduplication
17. Lead Scoring & Routing → Prioritization and assignment to sales officers
18. Lead Dossier Generation → Sales-ready summary created
19. Feedback Loop → Sales outcomes fed back for model improvement

3. Product Configuration

3.1 Product Object Schema

Each product in the `products` array has the following attributes:

Attribute	Type	Description
`code`	*string*	Product code (e.g., "HSD", "FO", "BITUMEN")
`name`	*string*	Full product name
`category`	*string*	"fuel" or "specialty"
`description`	*string*	Product overview and primary uses
`isActive`	*boolean*	Whether product is currently being tracked
`volumeProfile`	*string*	"low", "medium", "high", "very_high" - typical deal size
`typicalCustomerSize`	*array*	Customer segments: ["small", "medium", "large", "enterprise"]
`signalDetection`	*object*	All signal capture logic (see Section 4)
`industryMapping`	*object*	Industry classification and confidence levels
`operationalCues`	*array*	Keywords indicating operational use
`capacityIndicators`	*object*	Volume signals and minimum viable capacity
`useCases`	*array*	Primary use cases for the product
`geographicSignals`	*object*	High-demand regions and clusters
`inferenceLogic`	*object*	Confidence scoring rules (see Section 5)
`crossSellOpportunities`	*array*	Related products and triggers
`temporalPatterns`	*object*	Seasonal/cyclical demand patterns

4. Signal Detection Logic

4.1 Signal Categories

The `signalDetection` object uses a multi-layered approach to capture opportunities:

Layer	Purpose	Example	Impact
primaryKeywords	Direct product mentions	*Example: "furnace oil", "FO", "HSD", "bitumen"*	Highest confidence - explicit demand
secondaryKeywords	Operational/contextual terms	*Example: "boiler fuel", "DG set", "road construction"*	Medium confidence - implies need
contextualPhrases	Natural language patterns	*Example: "diesel requirement for DG", "bunker fuel procurement"*	Captures procurement language
negativeKeywords	Exclusion filters	*Example: "domestic", "household", "retail"*	Prevents false positives
tenderKeywords	Procurement-specific terms	*Example: "supply of", "tender", "annual requirement"*	Highest confidence - active buying
triggerSources	Source types to monitor	*Example: tender_portals, plant_expansion_news, capex_announcements*	Where to look for signals

4.2 Industry Mapping

Companies are classified into confidence tiers:

- **highConfidenceIndustries**: Industries with known, consistent product demand
  Example (Furnace Oil): Steel, Cement, Chemical, Paper, Sugar
  Base confidence: 0.40-0.55
- **mediumConfidenceIndustries**: Industries with occasional or lower-volume demand
  Example (Furnace Oil): Food Processing, Paint & Coatings
  Base confidence: 0.30-0.45
- **industrySpecificSignals**: Sector-specific keywords that boost confidence
  Example (Steel): "rolling mill", "reheating furnace", "heat treatment"
  Adds +0.10 to +0.20 to base confidence

5. Confidence Scoring System

5.1 Scoring Overview

⚠️ **CRITICAL FOR DEVELOPERS:** The confidence score determines lead routing and prioritization. Understanding this section is essential for system integration.

Confidence scores range from **0.30 to 0.95** on a decimal scale, where:

Score Range	Signal Strength	System Action
0.90 - 0.95	Explicit tender with volume/capacity specified	Immediate WhatsApp alert + Auto-assign
0.80 - 0.89	Strong signal (plant commissioning, tender mention)	Immediate email + Dashboard highlight
0.70 - 0.79	Good signal (expansion + operational cue)	Email digest + Dashboard
0.60 - 0.69	Moderate signal (industry + capacity indicator)	Daily digest + Dashboard
0.45 - 0.59	Weak signal (industry only with some cue)	Review queue - human evaluation needed
Below 0.45	Very weak signal (industry mention only)	Logged but not surfaced to sales

5.2 Base Confidence Rules

Each product defines `confidenceRules` that set the starting confidence based on signal type:

Example: Furnace Oil Confidence Rules

- `"explicitFOTenderWithVolume": 0.95`
- `"boilerInstallationWithCapacity": 0.90`

- `"furnaceInHighConfidenceIndustry": 0.85`
- `"plantExpansionWithBoilerMention": 0.80`
- `"steamRequirementWithIndustry": 0.70`
- `"highConfidenceIndustryWithFacility": 0.55`
- `"industryOnlySignal": 0.40`

5.3 Scoring Factors (Modifiers)

After determining base confidence, `scoringFactors` are applied as additive modifiers:

Factor	Condition	Impact
+0.15	Capacity/volume explicitly specified	High
+0.12	NHAI/PWD government project (for Bitumen)	High
+0.10	Multiple units (boilers/gensets/vessels)	Medium
+0.10	Urgent timeline or near commissioning	Medium
+0.10	Commissioning timeline within 6 months	Medium
+0.08	Environmental clearance obtained	Medium
+0.08	Competitor supplier mentioned	Medium
+0.05	Existing HPCL customer (other products)	Low
-0.05	Competitor strongly entrenched	Negative
-0.30	Coal-only mention (for FO)	Disqualifier
-0.40	Gas-based only (for FO)	Disqualifier

5.4 Final Score Calculation

Formula:

```
Final Confidence = BASE_CONFIDENCE + Σ(applicable_scoring_factors)
```

Constraints:

• Minimum score: 0.30 (below this = discarded)
• Maximum score: 0.95 (capped at this value)
• Rounding: 2 decimal places

5.5 Worked Example

Scenario: Steel company announces 20 TPH boiler installation for their new rolling mill

Step 1 - Signal Matching:

• Primary keyword: "boiler" ✓
• Industry: Steel (highConfidenceIndustry) ✓

• Operational cue: "rolling mill" ✓
• Capacity mentioned: "20 TPH" ✓

Step 2 - Base Confidence:
Match: "boilerInstallationWithCapacity" → **0.90**

Step 3 - Apply Scoring Factors:
• capacityVolumeSpecified: +0.15
• industrySpecificSignal (rolling mill): +0.10 (implicit in industry mapping)

Step 4 - Calculate:
`0.90 (base) + 0.15 (capacity) =` **1.05 → capped at 0.95**

Result: Confidence = 0.95 → Immediate WhatsApp alert + Auto-assign to sales officer

6. Validation & Quality Control

6.1 Quality Gates

The `verificationLayer` implements multiple quality checks before leads reach sales officers:

Gate	Threshold	Purpose	Type
Minimum Confidence Threshold	`0.60`	Leads below this are discarded	*Hard filter*
Human Review Threshold	`0.75`	Leads below this require manual review before assignment	*Quality gate*
Auto-Approve Threshold	`0.90`	Leads at/above this are auto-assigned without review	*Fast-track*
Source Credibility Check	`>0.6`	Source must have credibility score above threshold	*Quality filter*
Signal Freshness	`<90 days`	Signal must be less than 90 days old	*Recency filter*
Company Identity Resolved	`Required`	Company must be successfully normalized	*Entity resolution*
Minimum Viable Capacity	`Product-specific`	Must meet minimum volume threshold per product	*Size filter*

6.2 Deduplication Rules

The system prevents duplicate leads using these rules:

- **Same Company + Same Product** (Within 30 days)
  Prevents sending duplicate product leads for same company
- **Same Signal Source + Same Company** (Within 7 days)
  Prevents duplicate detection from same news article/tender
- **Cross-Product Deduplication** (Combine into multi-product lead)
  If multiple products detected for same company, create unified lead

6.3 Decision Flow

Lead Processing Decision Tree:

```
1. Signal detected → Extract company and product signals

2. Calculate base confidence → Apply confidenceRules

3. Apply scoring factors → Add/subtract modifiers

4. Check minimum threshold → If < 0.60: DISCARD

5. Check deduplication → If duplicate: MERGE or SKIP

6. Check quality gates → Source credibility, freshness, capacity

7. Determine routing:
```

- • If ≥ 0.90: Auto-assign + Immediate alert
- • If ≥ 0.75 and < 0.90: Auto-assign + Email digest
- • If ≥ 0.60 and < 0.75: Human review queue

```
8. Generate lead dossier → Create sales-ready summary

9. Route to sales officer → Territory-based assignment

10. Log for feedback loop → Track outcome for model improvement
```

7. Alerting & Routing Logic

7.1 Alerting Thresholds

The `alertingRules` object defines how leads are surfaced to sales officers:

Score	Alert Type	Channels	Delivery	Priority
0.85+	High Confidence Immediate Alert	WhatsApp + Email + Dashboard	Real-time push	**Top priority**
0.65 - 0.84	Medium Confidence Digest Alert	Email + Dashboard	Hourly/Daily digest	Standard
0.60 - 0.64	Low Confidence Review Queue	Dashboard only	Manual review needed	Review

7.2 Urgency Factors

Urgency can override standard alerting based on these factors:

- **Tender Deadline Near**: Within 15 days
  → *Escalate to immediate alert even if confidence = 0.75*
- **Competitor Activity**: Competitor mentioned in signal
  → *Add urgency flag to alert*
- **Large Opportunity**: Estimated value > threshold
  → *Priority routing to senior sales officer*
- **Existing Customer**: Company in HPCL database
  → *Route to account manager immediately*

8. Implementation Guide for Developers

8.1 JSON Structure Access

Loading the Dataset (Python):

```
import json

# Load the dataset
with open('hpcl_product_intelligence_complete.json', 'r') as f:
    dataset = json.load(f)

# Access product configurations
products = dataset['products']
furnace_oil = next(p for p in products if p['code'] == 'FO')

# Access signal detection keywords
primary_keywords = furnace_oil['signalDetection']['primaryKeywords']
# ['furnace oil', 'FO', 'heavy fuel oil', 'HFO', 'LSHS']

# Access confidence rules
confidence_rules = furnace_oil['inferenceLogic']['confidenceRules']
# {'explicitFOTenderWithVolume': 0.95, 'boilerInstallationWithCapacity': 0.90, ...}

# Access validation thresholds
min_confidence = dataset['verificationLayer']['minimumConfidenceThreshold']
# 0.60
```

8.2 Signal Matching Logic

Example Signal Matching Function (Pseudocode):

```
function matchSignalsToProduct(signalText, productConfig):
    score = 0.0
    matchedKeywords = []

    # Check primary keywords (highest weight)
    for keyword in productConfig.signalDetection.primaryKeywords:
        if keyword.lower() in signalText.lower():
            matchedKeywords.append(('primary', keyword))
            score += 0.3

    # Check secondary keywords
    for keyword in productConfig.signalDetection.secondaryKeywords:
        if keyword.lower() in signalText.lower():
            matchedKeywords.append(('secondary', keyword))
            score += 0.15

    # Check negative keywords (disqualify)
    for keyword in productConfig.signalDetection.negativeKeywords:
        if keyword.lower() in signalText.lower():
            return None  # Signal rejected

    # Check tender keywords (boost score)
    for keyword in productConfig.signalDetection.tenderKeywords:
        if keyword.lower() in signalText.lower():
            matchedKeywords.append(('tender', keyword))
            score += 0.4

    return {
        'matchScore': min(score, 1.0),
        'matchedKeywords': matchedKeywords,
        'productCode': productConfig.code
    }
```

8.3 Confidence Calculation

Example Confidence Scoring Function:

```
function calculateConfidence(signal, productConfig, companyProfile):
    # Step 1: Determine base confidence from rules
    baseConfidence = determineBaseConfidence(signal, productConfig)

    # Step 2: Apply scoring factors
    finalConfidence = baseConfidence
    scoringFactors = productConfig.inferenceLogic.scoringFactors

    # Check each factor
    if hasCapacityMention(signal):
        finalConfidence += scoringFactors.capacityMentioned  # +0.15

    if hasMultipleUnits(signal):
        finalConfidence += scoringFactors.multipleGensets  # +0.10
```

```
    if hasUrgencyIndicators(signal):
        finalConfidence += scoringFactors.urgencyIndicators  # +0.10

    if isExistingCustomer(companyProfile):
        finalConfidence += scoringFactors.existingHPCLCustomer  # +0.05

    if competitorMentioned(signal):
        finalConfidence += scoringFactors.competitorMentioned  # +0.08

    # Step 3: Apply constraints
    finalConfidence = max(0.30, min(finalConfidence, 0.95))
    finalConfidence = round(finalConfidence, 2)

    return {
        'baseConfidence': baseConfidence,
        'finalConfidence': finalConfidence,
        'appliedFactors': getAppliedFactors(),
        'reasonCodes': generateReasonCodes(signal, productConfig)
    }

function determineBaseConfidence(signal, productConfig):
    rules = productConfig.inferenceLogic.confidenceRules

    # Check in priority order (highest confidence first)
    if isTenderWithVolume(signal):
        return rules.explicitTenderWithVolume  # 0.95

    if isPlantCommissioning(signal) AND hasCapacity(signal):
        return rules.plantCommissioningWithCapacity  # 0.90

    if isIndustryMatch(signal) AND hasFacilityMention(signal):
        return rules.industryWithFacilityMention  # 0.70-0.85

    if isIndustryMatchOnly(signal):
        return rules.industryOnlySignal  # 0.35-0.45

    return 0.30  # Minimum baseline
```

8.4 Validation Workflow

Lead Validation Pipeline:

```
function validateLead(lead, dataset):
    verificationLayer = dataset.verificationLayer

    # Gate 1: Minimum confidence
    if lead.confidence < verificationLayer.minimumConfidenceThreshold:
        return {'status': 'REJECTED', 'reason': 'Below minimum confidence (0.60)'}

    # Gate 2: Deduplication
    if isDuplicate(lead, verificationLayer.deduplicationRules):
        return {'status': 'DUPLICATE', 'reason': 'Same company/product within 30 days'}

    # Gate 3: Source credibility
    if lead.sourceCredibility < 0.6:
        return {'status': 'REJECTED', 'reason': 'Source credibility too low'}

    # Gate 4: Signal freshness
    if lead.signalAge > 90:  # days
```

```
        return {'status': 'REJECTED', 'reason': 'Signal older than 90 days'}

    # Gate 5: Company identity resolved
    if not lead.companyResolved:
        return {'status': 'PENDING', 'reason': 'Awaiting entity resolution'}

    # Gate 6: Minimum viable capacity
    if not meetsMinimumCapacity(lead):
        return {'status': 'REJECTED', 'reason': 'Below minimum viable capacity'}

    # Determine routing
    if lead.confidence >= 0.90:
        return {'status': 'APPROVED', 'routing': 'AUTO_ASSIGN', 'alert': 'IMMEDIATE'}
    elif lead.confidence >= 0.75:
        return {'status': 'APPROVED', 'routing': 'AUTO_ASSIGN', 'alert': 'DIGEST'}
    else:  # 0.60 - 0.74
        return {'status': 'APPROVED', 'routing': 'REVIEW_QUEUE', 'alert': 'DASHBOARD'}
```

8.5 Integration Checklist

- ☐ Load and parse the JSON dataset into your application
- ☐ Implement signal matching logic for all keyword layers
- ☐ Build confidence calculation engine with base rules + factors
- ☐ Create validation pipeline with all quality gates
- ☐ Implement deduplication logic (30-day and 7-day windows)
- ☐ Set up alerting channels (WhatsApp, Email, Dashboard)
- ☐ Configure territory-based routing for lead assignment
- ☐ Build lead dossier generation with reason codes
- ☐ Implement feedback loop to capture sales outcomes
- ☐ Set up monitoring and logging for all confidence scores
- ☐ Create admin interface for threshold adjustments
- ☐ Build analytics dashboard for lead performance tracking

9. Appendices

9.1 Complete Confidence Scale Reference

Score	Condition	Example
0.95	Explicit tender with volume AND capacity specified	*Tender: "Supply 500 MT FO monthly for 20 TPH boiler"*
0.90	Plant commissioning with equipment capacity	*News: "Steel plant commissioning 30 TPH boiler in Q2"*
0.85	Equipment installation in high-confidence industry	*Press release: "Cement plant installing rotary kiln"*
0.80	Facility expansion with product-relevant equipment	*Article: "Chemical plant expanding with new furnace"*
0.75	Tender mention without full volume details	*Tender: "Supply of furnace oil" (no volume specified)*
0.70	Industry match + strong operational cue	*Website: "Steel mill with rolling operations"*
0.65	Industry match + moderate operational cue	*Directory: "Sugar mill with cogeneration"*
0.60	Industry match + minimal cue	*Listing: "Paper manufacturing unit"*
0.55	High-confidence industry + facility mention	*Report: "Fertilizer plant in Gujarat"*
0.45	Medium-confidence industry + some cue	*News: "Food processing expansion"*
0.40	High-confidence industry mention only	*Directory: "Cement manufacturer"*
0.35	Medium-confidence industry only	*Listing: "Paint company"*
0.30	Weak signal, minimum threshold	*Mention: "Industrial operations"*

9.2 Product-Specific Minimum Capacities

Product	Minimum Viable Capacity	Rationale
HSD	`250 KVA genset OR 10 KL monthly`	Below this = residential/small commercial
LDO	`50 KVA genset OR 2 KL monthly`	Small capacity threshold
Furnace Oil	`5 TPH boiler OR 100 MT monthly`	Minimum industrial scale
Bitumen	`5 km road project OR 100 MT`	Viable road construction
Marine Bunker	`1000 MT monthly OR fleet of 3+ vessels`	Commercial shipping operations
Hexane	`50 TPD seed crushing OR 10 KL monthly`	Industrial extraction scale
Propylene	`1000 MT annual OR 50 MTPA polymer capacity`	Chemical/polymer production
JBO	`Active jute mill with regular production`	Operational mill
Solvent 1425	`5 KL monthly OR 1000 L/day paint capacity`	Manufacturing scale
MTO 2445	`5 KL monthly consumption`	Industrial usage
Sulphur	`500 MT annual OR 10,000 MTPA fertilizer`	Industrial scale fertilizer/chemical
SKO Non-PDS	`1 KL monthly consumption`	Commercial/industrial use

9.3 Cross-Product Opportunity Matrix

When certain signals are detected, evaluate multiple products:

- **Large Manufacturing Facility** → HSD + FO + LDO
  Multiple fuel needs

- **Fertilizer Plant** → Sulphur + FO + HSD
  Feedstock + thermal + power
- **Road Construction Project** → Bitumen + HSD
  Paving + equipment fuel
- **Port Operations** → Bunker + HSD
  Marine fuel + cargo handling
- **Oil Mill / Extraction Plant** → Hexane + HSD + FO
  Extraction + power + steam
- **Paint Manufacturer** → Solvent 1425 + MTO 2445
  Multiple solvent types
- **Chemical Plant** → Multiple specialty products
  Process-dependent

10. Support & Maintenance

10.1 Dataset Updates

This dataset should be reviewed and updated:

- Monthly: Review feedback from sales team on lead quality
- Quarterly: Analyze conversion rates and adjust confidence thresholds
- Semi-annually: Add new keywords based on missed opportunities
- Annually: Comprehensive review of all products and scoring rules
- Ad-hoc: When new product lines are introduced or major market changes occur

10.2 Feedback Loop

Continuous improvement requires tracking:

- Lead acceptance rate (% of leads acted upon by sales)
- Lead conversion rate (% of leads resulting in orders)
- False positive rate (% of leads rejected as irrelevant)
- Time from signal to sales action
- Revenue per lead (to validate scoring accuracy)
- Keyword effectiveness (which keywords correlate with conversions)