Making India's data AI-ready

Every move, every spend, every signal — captured, cleaned, and structured for the age of artificial intelligence

Akhilesh Tilotia
Co-founder, Thurro
March 2026

Thurro
Structured Intelligence

Making India's data AI-ready

The data explosion: Millions of real-time signals daily across payments, production, prices, and logistics — most arriving as PDFs, HTML tables, or locked portals

The barriers: CAPTCHAs, format changes, schema drift, missing metadata, zero API standards — 80% of analyst time goes into cleaning data, not analysing it

What government is already doing: MoSPI's MCP server, PM Gatishakti, ONDC are first steps toward machine-readable public data — but much more is needed

The cost of inaction: Models on stale data, cross-source joins that fail, dashboards that break — messy data is a structural barrier to AI in India, not just a nuisance

What AI-ready looks like: Standardised APIs, documented schemas, revision history, universal data models, complete metadata, and self-healing pipelines

Three initiatives India should pursue

- **Open up transformative datasets** — GST microdata, MCA company financials, NPCI payment trends, survey microdata, and administrative records
- **Build a sovereign SLM for Indian finance** — a domain-specific model trained on local regulations, company filings, and Indian reporting idiosyncrasies
- **Make Census 2027 AI-ready** — structured schemas, privacy architecture, and API-first data access designed from day one

India generates millions of real-time data signals every day

15M+ data points daily at Thurro, drawn from **800+ sources** across Indian capital markets

Payments and spending

UPI volumes, credit/debit card spends, NPCI aggregates, festival spending, digital wallets, BBPS bill payments...

Production and industry

IIP, auto production, steel output, core sector, manufacturing PMI, cement dispatches, coal output...

Movement and logistics

FASTag tolls, port cargo, freight index, aviation passenger flows, railway freight volumes, e-way bills...

Consumption and retail

Store locations, FMCG data, QSR sales, mall traffic, SKU-level pricing from Blinkit, Amazon, Dmart, et al, hotel prices, airfare...

Investment and capital

Demat accounts, SIP inflows, bulk/block deals, gold ETFs, FII/DII flows, insider trades, MF portfolios...

Prices and inflation

CPI and WPI baskets, fuel prices, mandi prices, rent indices, retail price tracking, AGMARKNET data...

Why Indian alternate data is so hard to work with

Collection barriers

- CAPTCHAs, IP throttling, and bot detection impede full browser automation
- Rate-limiting, login walls, and URL changes block automated access to ostensibly public data
- Portal redesigns break every parser overnight — HTML tables turn into PDFs with no advance warning
- Data due on the 12th appears on the 18th — or is silently delayed by two months with no explanation

Quality problems

- Historical revisions — Many figures are revised months after initial release; need to keep updating the data
- Missing data — states report at different cadences, fields are blank, aggregates do not match sums
- Schema drift — NIC codes update, indices get reweighted, and base years change
- No metadata standards — limited data dictionaries, few machine-readable schema, every portal is different

The cost

~80%

of analyst time goes into cleaning, not analysis

Weeks

to onboard each new source from discovery to production

Zero

Indian alt-data sources that ship with API docs or schemas

Build infrastructure that makes Indian data AI-ready

Today	AI-ready
PDFs, HTML, Excel, images No format consistency	**JSON, Parquet, CSV** Documented schemas
Custom scrapers per source Fragile, break overnight	**Standardised APIs** Versioned endpoints, rate limits
Revisions overwrite silently No version control	**Full revision history** Every data point timestamped
No metadata or schemas Each portal is its own world	**Universal data models** Across sources and codes
No provenance or audit trail Who published what, when?	**Complete metadata** Source, frequency, methodology
Emergency 2 AM fixes Schema breaks, no fallback	**Self-healing pipelines** Anomaly detection, alerting

Leverage government initiatives that are making data AI-ready now

MoSPI MCP server — launched Feb 6, 2026

The Ministry of Statistics and Programme Implementation's Model Context Protocol server on the eSankhyiki portal lets AI tools query official statistics directly — no file downloads, no preprocessing. Seven datasets at launch: Periodic Labour Force Survey, Consumer Price Index, Annual Survey of Industries, Index of Industrial Production, National Accounts Statistics, Wholesale Price Index, and Energy Statistics

PM Gatishakti portal

Integrated infrastructure planning across 16 ministries. Opens spatial and logistics data for AI-driven physical planning, site selection, and supply chain optimisation

Open Network for Digital Commerce (ONDC)

Open digital commerce protocols across 630+ cities. Decoupled transactional data streams for payments and discovery, enabling granular demand-side signals

Open up six datasets that can transform analysis across sectors

GST revenue microdata

Goods and Services Tax collection by sector and state — identify growing clusters, map economic density, and benchmark state-level industrial performance

NPCI payment trends

National Payments Corporation of India data on merchant-level and peer-to-peer UPI trends by industry — real-time consumer spending signals at sectoral granularity

Survey microdata from MoSPI

Unit-level data from Periodic Labour Force Survey, consumer expenditure, and health surveys — granular demographic and labour market modelling

MCA company financials

Company-wise annual financials, director changes, charge filings, and incorporation data across all registered entities — sectoral health, credit risk, and corporate governance

PM Gatishakti spatial data

Integrated geospatial data across rail, road, port, and logistics infrastructure — physical planning, site selection, and supply chain analytics

Administrative and public records

Hospital admissions, disease incidence, insurance claims, medicines procured; EPFO enrolment; power consumption at transformer level

Build a sovereign SLM for Indian finance and economy

General-purpose large language models do not understand Indian regulatory filings, company naming conventions, vernacular financial reporting, or the idiosyncrasies of how Indian data is structured. A sovereign SLM trained on Indian financial and economic data can outperform generic models at a fraction of the cost — and keep sensitive financial intelligence within India's borders.

Local data context

Indian CPI uses a different basket and methodology from the US. IIP classification codes, GST filing formats, MCA annual return schemas — none of these exist in any global training corpus.

Regulatory idiosyncrasies

SEBI circulars, RBI master directions, IRDAI guidelines, TRAI regulations — thousands of pages of domain-specific rules that shape how markets, banks, and insurers operate.

Company and reporting quirks

Promoter pledging, related-party transactions, Indian GAAP vs Ind-AS transitions, group structures with dozens of subsidiaries — patterns only visible in Indian filings.

Sovereign and secure

Financial intelligence should not depend on models hosted abroad. A sovereign SLM trained on Indian data and hosted on Indian infrastructure ensures data sovereignty.

Make Census 2027 AI-ready — not just digital

India's last completed census was in 2011. Census 2027 will be fully digital — but digital is not the same as AI-ready. The difference: structured schemas, standardised classification codes, real-time validation, and machine-readable output from day one. The incremental cost is 10–15% of the baseline. The cost of not doing it: another decade of partially usable demographic data.

Structured schemas

Standardised codes for occupation (National Classification of Occupations), industry (NIC), and geography (Local Government Directory) validated at data entry — not cleaned years later.

Privacy architecture

Differential privacy, tiered access from public aggregates to restricted microdata, statutory firewalls against function creep — built into the design.

API-first data access

Machine-readable outputs in open formats from launch. No waiting years for cleaned tables. Census as live national data infrastructure.

India has the data.
Now make it AI-ready.

akhilesh@thurro.com

Chennai: 326, DBS Center, 31A Cathedral Garden Road, Chennai 600034
Mumbai: BHive, 3rd Floor, Inspire, BKC, Bandra East, Mumbai 400051

Thurro
Structured Intelligence