Alright — based on your expanded Threat Intelligence Platform phase map and the fact that you have a **Data/AI Engineer, Backend Engineer, Flutter Developer, and Security/Network Engineers** in the team, the **Data Engineer’s** role would focus on building, optimizing, and maintaining the **data backbone** of the platform so the rest of the team can consume and work with clean, enriched, and well-structured threat intelligence.

Here’s the breakdown of **Data Engineer job responsibilities** phase by phase:

**Phase 0 – Platform Architecture & Infrastructure Setup**

* Collaborate with Backend & Security Engineers to design **data architecture** (data lake, warehouse, hot/cold storage strategy).
* Select and configure **data pipelines** (Airflow, Prefect, Kafka, etc.).
* Set up **monitoring & logging** for all ETL/ELT jobs.

**Phase 1 – Data Acquisition & Ingestion**

* Build and maintain **multi-source ingestion pipelines** for STIX/TAXII, MISP, OTX, CSV/XML feeds, APIs, and streaming data.
* Implement **source validation** (integrity checks, trust scoring).
* Design and automate **real-time and batch ingestion** workflows.
* Monitor ingestion performance and **handle failed jobs/retries**.
* Store raw feeds in a **structured format** for downstream processing.

**Phase 2 – Data Normalization & Storage**

* Develop **parsers & transformation scripts** to standardize IOCs and threat data (IP formats, timestamp normalization, etc.).
* Implement **schema mapping** to industry standards (STIX 2.1, MISP schema).
* Create and maintain **optimized database schemas** for fast search and correlation.
* Manage **hot storage** (Elasticsearch/Redis) and **cold storage** (S3, PostgreSQL).
* Automate **data lifecycle policies** (archiving, deletion).

**Phase 3 – Threat Data Management**

* Implement **data enrichment jobs** (WHOIS, GeoIP, ASN lookups, Passive DNS, VirusTotal integrations).
* Build **tagging & classification automation** for IOCs.
* Maintain **version control of threat data** and historical tracking.
* Ensure all data meets **TLP classification tagging rules**.

**Phase 4 – Analysis & Correlation**

* Prepare **query-optimized datasets** for correlation algorithms and ML models.
* Build **ETL processes for threat scoring** and correlation workflows.
* Assist AI Engineer in **feeding cleaned/enriched datasets into LLMs** or ML models.
* Maintain historical **time-series datasets** for predictive analytics.

**Phase 5 – Visualization & Reporting**

* Build **data APIs** or Elasticsearch queries to feed dashboards.
* Ensure data structures are **optimized for backend and Flutter app queries**.
* Automate **report data extraction & formatting** for PDF/HTML/CSV outputs.

**Phase 6 – Security Operations Integration**

* Enable **real-time data streaming** to SIEM/SOAR connectors (Splunk, QRadar).
* Provide **data endpoints** for security playbooks.
* Maintain APIs that expose clean data for external SOC tools.

**Phase 7 – Governance, Compliance & Security**

* Implement **audit logging for data pipelines**.
* Work with Security Engineers to ensure **GDPR/CCPA compliance** in stored data.
* Manage **data retention policies** and secure deletion processes.
* Handle **secrets management** for API keys and integration credentials.

**Phase 8 – AI & Automation (if included)**

* Collaborate with AI Engineer to **prepare training datasets** from threat feeds.
* Automate **LLM preprocessing** (cleaning, summarizing, and formatting data before it hits the model).
* Maintain feedback loops for **model improvement based on real-world performance**.

**Expected Outcomes from Data Engineer’s Work**

* **Reliable, fast, and scalable** ingestion pipelines that keep the platform up-to-date with threat intelligence in real time.
* **Clean, enriched, and well-structured datasets** ready for search, analysis, AI models, and visualization.
* **Highly available data APIs** that power the backend, Flutter app, and external integrations.
* **Auditable and compliant** data flows meeting security and legal requirements.
* **Automation** in enrichment, normalization, and correlation that reduces analyst workload.

**UPDATED PLAN:**  
  
**Data Engineer – Job Responsibilities in the TIP Pipeline**

**1. Threat Intelligence Data Ingestion & ETL**

* Design, build, and maintain **data ingestion pipelines** for multiple threat intelligence sources:
  + Structured feeds (STIX/TAXII, CSV, JSON APIs)
  + Semi-structured/unstructured OSINT sources (blogs, forums, paste sites)
  + Dark web monitoring sources (via scraping or APIs)
* Handle **real-time streaming ingestion** from security tools (e.g., SIEM logs, IDS/IPS alerts).
* Implement **ETL/ELT** workflows to transform and normalize diverse feed formats into a unified schema.

**2. Data Storage & Management**

* Architect and maintain **centralized threat intelligence databases** (e.g., PostgreSQL, Elasticsearch, or graph DBs for relationships between IoCs).
* Manage **data partitioning, indexing, and lifecycle policies** to keep queries fast and storage efficient.
* Create and maintain **data lake** or **warehouse** for historical analysis and AI training datasets.

**3. Data Cleaning & Normalization**

* Develop scripts to **deduplicate, validate, and enrich** incoming data.
* Map and align data to **standard taxonomies** (MITRE ATT&CK, CVE, CWE, CAPEC).
* Ensure data quality through **automated validation checks** and error handling.

**4. Collaboration with AI Engineers**

* Provide AI engineers with **preprocessed, labeled, and structured datasets** ready for model training.
* Implement **feature engineering** pipelines to extract relevant features from raw threat data.
* Store and manage **versioned datasets** for reproducible AI model training.

**5. Collaboration with Backend Engineers**

* Design database schemas and APIs for **fast retrieval** of threat intelligence by backend services.
* Optimize queries and storage strategies to meet **real-time threat detection requirements**.
* Support backend developers in **building search, correlation, and alerting features**.

**6. Data Security & Compliance**

* Work with Security/Network Engineers to:
  + Secure data pipelines against tampering or unauthorized access.
  + Implement encryption at rest and in transit.
  + Ensure compliance with **data handling regulations** (GDPR, CCPA if applicable).

**7. Monitoring & Optimization**

* Implement **data pipeline monitoring** (e.g., Prefect, Airflow, Dagster) for job success/failure tracking.
* Tune performance for **large-scale ingestion** and high-volume feeds.
* Automate scaling for peak load during major cyber events.

✅ **Expected Outcomes**:

* Reliable, scalable, and secure pipelines delivering high-quality threat data to the AI, backend, and frontend teams.
* Unified data model supporting cross-source correlation and advanced analytics.
* Well-maintained historical dataset enabling accurate AI predictions and threat trend analysis.