

Data Flow Document

Data Flow Document

“ **Project:** {{PROJECT_NAME}} **Version:** {{VERSION}} **Date:** {{DATE}}
Author: {{AUTHOR}} **Status:** Draft | In Review | Approved **Reviewers:**
{{REVIEWERS}} **Classification:** Public | Internal | Confidential | Restricted

Document History

Version	Date	Author	Changes
0.1	{{DATE}}	{{AUTHOR}}	Initial draft

1. Data Flow Overview

System: {{SYSTEM_NAME}} **Data Owner:** {{DATA_OWNER_ROLE}} **DPO Contact:**
{{DPO_EMAIL}}

Overview: {{DESCRIBE_WHAT_DATA_FLOWS_THROUGH_SYSTEM}}

High-Level Data Flow

```
flowchart LR
    subgraph Inputs["Data Sources / Ingestion"]
        U[Users – Web/App]
        API_IN[External API]
        IMPORT[Bulk Import]
        WEBHOOK[Webhooks]
    end

    subgraph Processing["Processing Layer"]
```

```
    VAL[Validation & Sanitization]
    TRANS[Business Logic / Transformation]
    ENRICH[Data Enrichment]
end

subgraph Storage["Storage Layer"]
    DB[(Primary DB\nPostgreSQL)]
    CACHE[(Cache\nRedis)]
    BLOB[Object Storage\nS3/Blob]
    SEARCH[Search Index\nElasticsearch]
    DW[Data Warehouse\n{{DW_TECH}}]
end

subgraph Outputs["Data Consumers / Egress"]
    API_OUT[REST API]
    REPORTS[Reports / Analytics]
    EXPORT[Data Export]
    THIRD[Third-party Integrations]
    EMAIL[Email / Notifications]
end

U & API_IN & IMPORT & WEBHOOK --> VAL
VAL --> TRANS
TRANS --> ENRICH
ENRICH --> DB & BLOB
DB --> CACHE & SEARCH
DB --> DW
DB --> API_OUT & REPORTS & EXPORT
DW --> REPORTS
ENRICH --> THIRD
DB --> EMAIL
```

2. Data Sources & Ingestion

Source	Type	Protocol	Volume (est.)	Format	PII?	Validation
--------	------	----------	---------------	--------	------	------------

Web application users	Real-time	HTTPS POST	{{REQ_PER_DAY}} req/day	JSON	YES	Schema + business rules
Mobile app users	Real-time	HTTPS POST	{{REQ_PER_DAY}} req/day	JSON	YES	Schema + business rules
{{EXTERNAL_SYSTEM}} API	Real-time	Webhooks	{{EVENTS_PER_DAY}} events/day	JSON	{{YES/NO}}	HMAC signature + schema
CSV bulk import	Batch	File upload	{{IMPORTS_PER_DAY}} files/day	CSV	{{YES/NO}}	Column mapping + row validation
{{THIRD_PARTY_API}}	Polling	REST/HTTPS	{{CALLS_PER_HOUR}}/hour	JSON	{{YES/NO}}	Response schema validation

Ingestion Error Handling

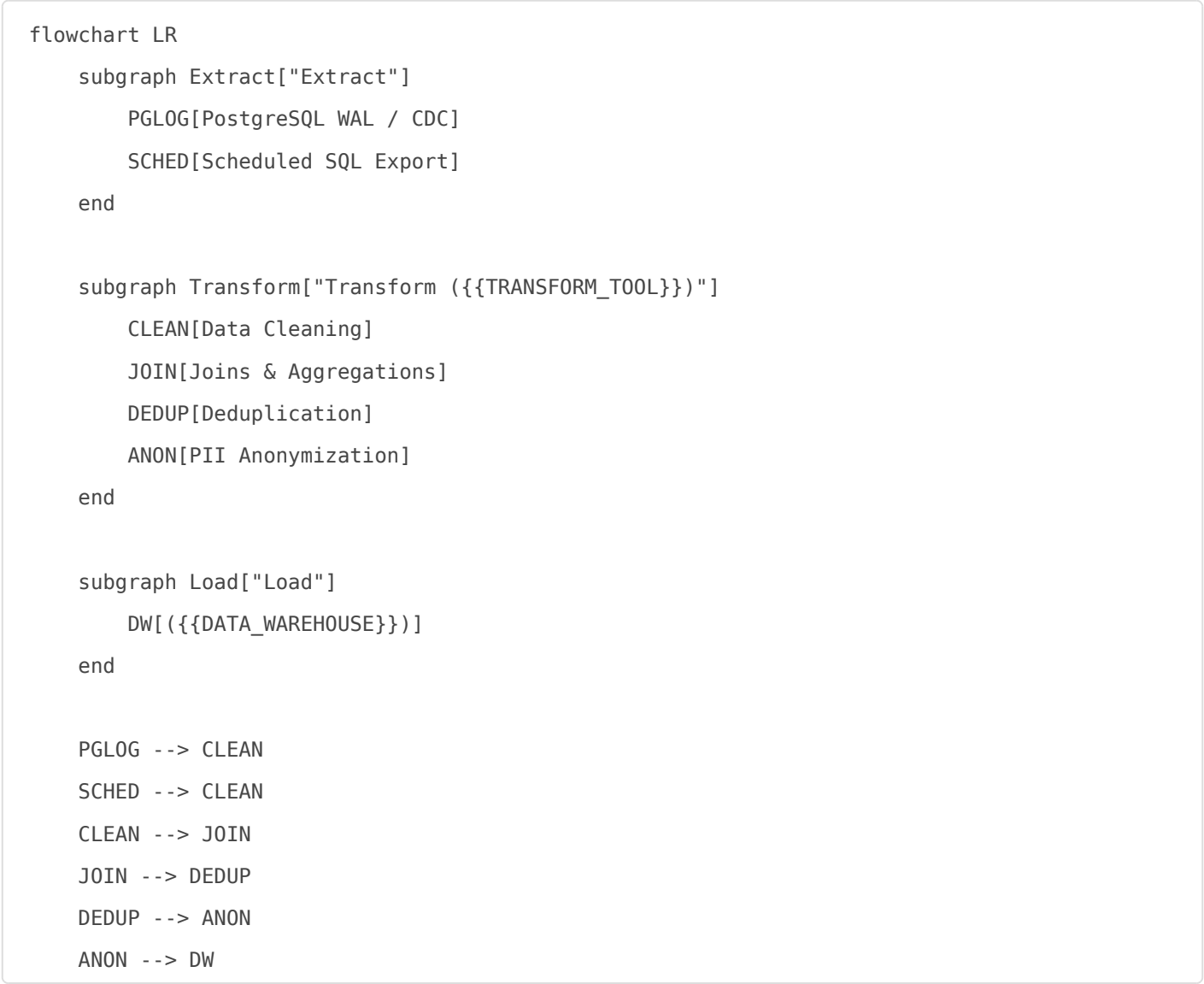
Error Type	Action	Notification
Schema validation failure	Reject with error details	Return 400 to caller
Duplicate record	Upsert (prefer existing) or reject	Log, return 409
PII fields contain unexpected data	Quarantine + alert	Slack #{{CHANNEL}}
Import file corrupted	Reject entire file	Email uploader + error report

3. Data Transformations

3.1 Ingestion Transformations (before storage)

Step	Input	Transformation	Output	Notes
1. Sanitization	Raw user input	Strip HTML, trim whitespace	Clean strings	Prevents XSS
2. Normalization	{{FIELD}}	Lowercase + trim	Normalized {{FIELD}}	e.g., email normalization
3. Enrichment	User IP	GeoIP lookup	{country, region, city}	Third-party API call
4. PII masking	{{PII_FIELD}}	Hash / mask for logs	Masked value	Never log raw PII
5. Encryption	Sensitive fields	AES-256-GCM	Encrypted blob	At application layer

3.2 ETL Pipeline (to Data Warehouse)



Pipeline schedule: `{{PIPELINE_SCHEDULE}}` (e.g., hourly incremental, daily full) **Latency:** Source to DW within `{{MAX_LATENCY}}` **Tool:** `{{ETL_TOOL}}` (e.g., dbt, Airbyte, custom)

4. Data Storage

Storage System	Technology	Purpose	Data Classification	Encryption at Rest
Primary Database	<code>{{DB_TECH}}</code> <code>{{VERSION}}</code>	Transactional data	Confidential	AES-256 (<code>{{KEY_MGMT}}</code>)
Cache	Redis <code>{{VERSION}}</code>	Hot data, sessions	Internal	AES-256
Object Storage	<code>{{S3_COMPATIBLE}}</code>	Files, documents, media	<code>{{CLASSIFICATION}}</code>	SSE-S3 / SSE-KMS

Storage System	Technology	Purpose	Data Classification	Encryption at Rest
Search Index	Elasticsearch	Full-text search	Internal	TLS + at-rest encryption
Data Warehouse	{{DW}}	Analytics, reporting	Anonymized	{{DW_ENCRYPTION}}
Backup Storage	{{BACKUP_TECH}}	Disaster recovery	Restricted	AES-256
Audit Logs	{{LOG_STORAGE}}	Compliance / audit trail	Restricted	Immutable, encrypted

5. Data Access Patterns

5.1 Read Patterns

Consumer	Data Accessed	Frequency	Access Method	Caching
Web application	User profile, settings	Per request	REST API	Redis 5min TTL
Web application	{{ENTITY}} list	Per page load	REST API (paginated)	CDN + Redis
Reporting service	Aggregated metrics	Every 1h	DW query	Materialized views
Admin dashboard	Raw records	On demand	REST API (admin)	No cache
External partner	{{SUBSET_OF_DATA}}	{{FREQUENCY}}	REST API (scoped JWT)	{{CACHING}}

5.2 Write Patterns

Writer	Data Written	Frequency	Write Method	Consistency
User actions (web)	CRUD operations	Per user action	REST API	Strong (synchronous)
Background worker	Aggregates, computed fields	Every {{INTERVAL}}	Direct DB write	Eventual
Import process	Bulk records	{{FREQUENCY}}	Batch insert	Strong (per batch)
Event consumer	Denormalized cache	On event	Direct DB write	Eventual

6. Data Retention & Archival

Data Category	Retention Period	Legal Basis	Action at Expiry	Automated?
---------------	------------------	-------------	------------------	------------

User account data	Duration of relationship + {{N}} years	Contract	Soft delete → anonymize	Automated (nightly job)
Transaction records	{{N}} years	Legal obligation ({{REGULATION}})	Archive to cold storage	Automated
Audit logs	{{N}} years	Legitimate interest (security)	Delete	Automated
Session tokens	{{N}} hours/days	Technical necessity	Auto-expire via TTL	Yes (Redis TTL)
Marketing consent	Until withdrawn	Consent	Delete within {{N}} days of withdrawal	Manual + automated
Analytics data	{{N}} years (anonymized)	Legitimate interest	Delete	Automated
Backup files	{{N}} days	Business continuity	Overwrite (rolling)	Automated
Error logs	{{N}} days	Legitimate interest	Delete	Automated

Retention schedule job: `retention-policy.job.ts` — runs daily at {{TIME}} UTC **Archival target:** {{COLD_STORAGE_LOCATION}}

7. Data Quality Rules

7.1 Validation Rules

Field	Rule	Error Action	Severity
<code>email</code>	Valid RFC 5322 format	Reject	CRITICAL
<code>phone</code>	E.164 format	Reject	HIGH
<code>{{DATE_FIELD}}</code>	Not in future	Reject	HIGH
<code>{{AMOUNT_FIELD}}</code>	>= 0	Reject	CRITICAL
<code>{{FK_FIELD}}</code>	References existing record	Reject	CRITICAL
<code>{{TEXT_FIELD}}</code>	Max {{N}} characters	Reject	MEDIUM

7.2 Data Quality Metrics

Metric	Target	Current	Alert Threshold
Null rate on required fields	0%	{{CURRENT}}	> 0.1%
Duplicate rate	< 0.01%	{{CURRENT}}	> 0.1%

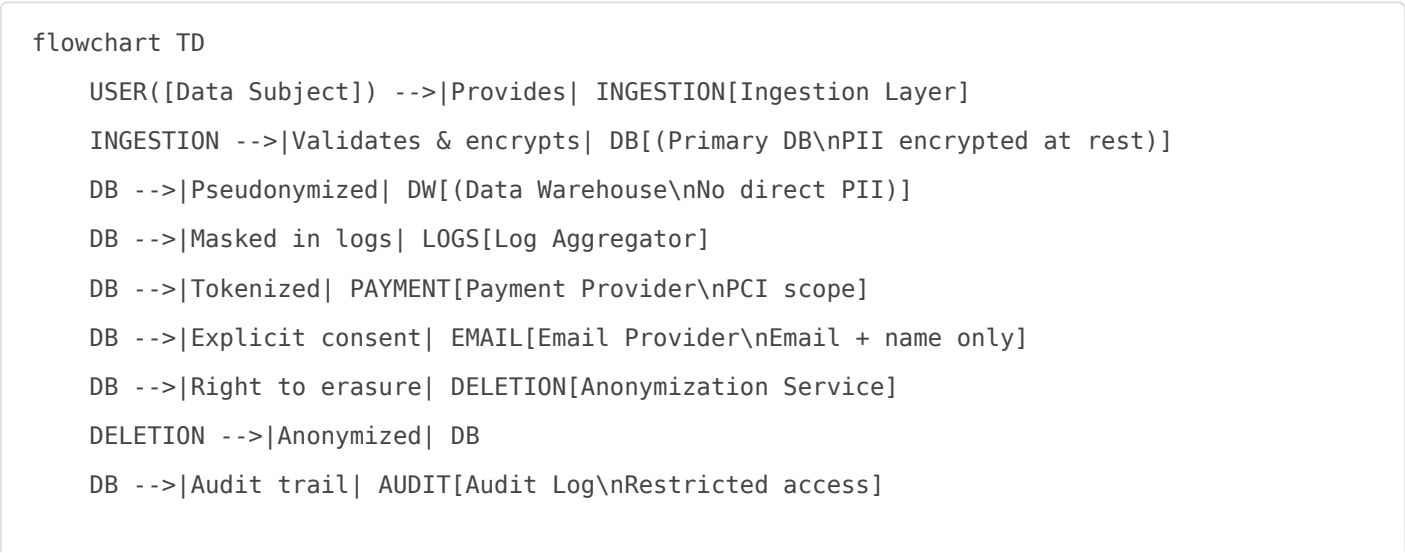
Metric	Target	Current	Alert Threshold
Schema validation pass rate	> 99.9%	{{CURRENT}}	< 99%
ETL pipeline success rate	> 99.5%	{{CURRENT}}	< 98%

8. PII Data Flow Mapping

8.1 PII Inventory

PII Category	Fields	Storage Location	Encrypted?	Access Controls	Lawful Basis
Contact info	email , phone , address	Primary DB, Email system	Yes	Role-based (user self + admin)	Contract
Identity	full_name , date_of_birth	Primary DB	Yes (field-level)	Role-based	Contract
Financial	{{PAYMENT_FIELD}}	{{PAYMENT_PROVIDER}} (tokenized)	Tokenized	PCI scope only	Contract
Behavioral	login_history , click_events	Analytics DB	No (anonymized)	Admin only	Legitimate interest
Location	ip_address (→ geo)	Logs (masked)	N/A	Admin only	Legitimate interest
Device	user_agent , device_id	Analytics DB	No	Admin only	Legitimate interest

8.2 PII Flow Diagram



style DB fill:#ffcccc

style DW fill:#ccffcc

style LOGS fill:#ffffcc

style AUDIT fill:#ffcccc

9. Cross-Border Data Transfer

Transfer	From	To	Data Category	Mechanism	DPA/SCCs?
{{TRANSFER_1}}	EU ({{COUNTRY}})	US ({{PROVIDER}})	{{DATA_CATEGORY}}	Standard Contractual Clauses (SCCs)	Yes — signed {{DATE}}
{{TRANSFER_2}}	EU	{{COUNTRY}}	{{DATA_CATEGORY}}	Adequacy decision	N/A
{{TRANSFER_3}}	EU	{{COUNTRY}}	{{DATA_CATEGORY}}	Binding Corporate Rules	{{YES/NO}}

Third-party processors with data access:

Processor	Service	Data Accessed	DPA Signed	Location
{{PROCESSOR_1}}	{{SERVICE}}	{{DATA}}	Yes	{{LOCATION}}
{{PROCESSOR_2}}	{{SERVICE}}	{{DATA}}	Yes	{{LOCATION}}

10. Data Lineage Tracking

Lineage tool: {{LINEAGE_TOOL}} (e.g., Apache Atlas, DataHub, custom) **Coverage:** Primary DB + DW

Lineage Events Captured

```
{
  "eventType": "DATA_WRITE",
  "timestamp": "IS08601",
  "actor": "system/user-id",
  "action": "CREATE | UPDATE | DELETE | EXPORT | IMPORT",
  "resource": {
    "type": "{{ENTITY}}",
```

```
    "id": "UUID"
  },
  "fields_modified": ["{{field1}}", "{{field2}}"],
  "sourceSystem": "{{SOURCE}}",
  "traceId": "UUID"
}
```

11. Backup & Recovery for Data

Storage	Backup Method	Frequency	Retention	RTO	RPO	Test Frequency
Primary DB	Continuous WAL archiving + snapshots	Continuous / Daily	30 days	1h	5min	Monthly
Object Storage	Cross-region replication	Continuous	30 days	4h	1h	Quarterly
Data Warehouse	Snapshot	Daily	14 days	8h	24h	Quarterly
Redis Cache	RDB snapshots	Every 15min	24h	15min	15min	Monthly

Last backup test: {{DATE}} — Result: {{PASS/FAIL}} **Recovery runbook:** {{LINK_TO_RUNBOOK}}

Approval

Role	Name	Date	Signature
Author			
Data Owner			
DPO / Privacy			
Security			
Tech Lead			