Stripe Business Case

Data Pipelines for AI

Context

- **Stripe**: growing online payment processing platform
- Increasing complexity of data management across platforms
- Necessity to refactor the data infrastructure and pipelines
- Integration of the infrastructure with data sources and consumers

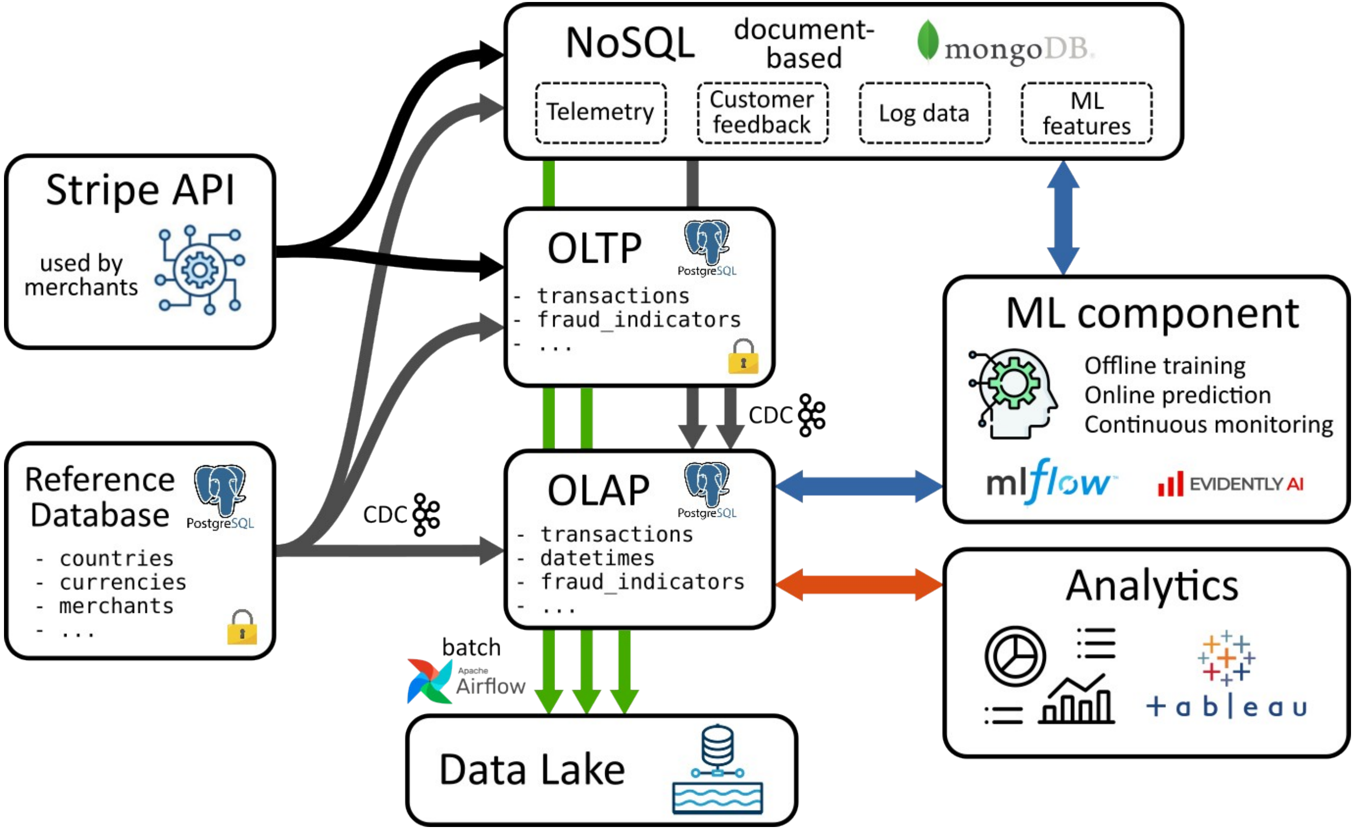
Context

- **Stripe**: growing online payment processing platform
- Increasing complexity of data management across platforms
- Necessity to refactor the data infrastructure and pipelines
- Integration of the infrastructure with data sources and consumers

Tasks

- OLTP data model
- OLAP data model
- NoSQL data model
- Data integration architecture
- Security and compliance plan
- Machine learning integration
- Example SQL and NoSQL queries

Pipeline architecture



Reference Database

- Slowly changing/static data

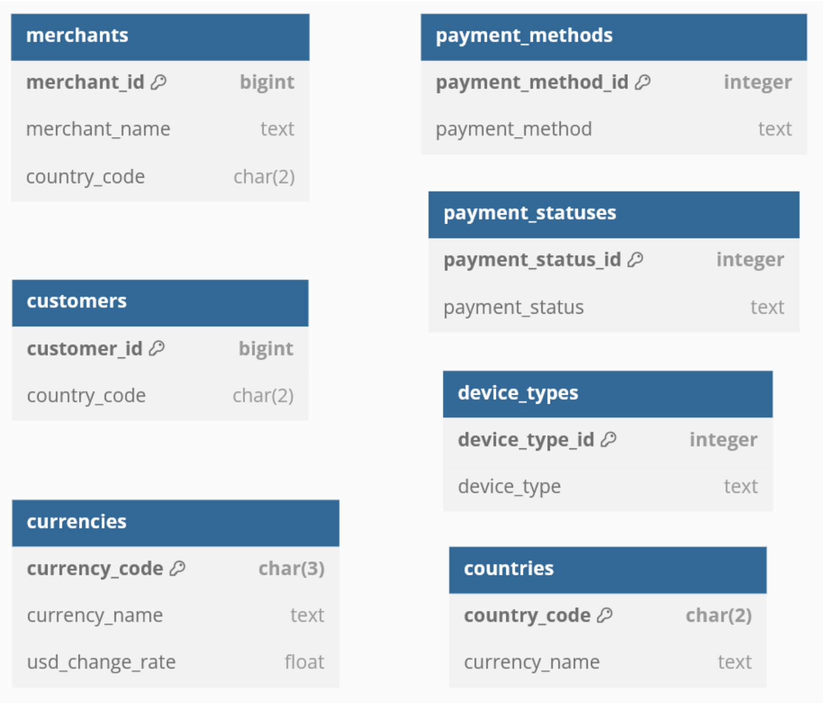
  e.g., countries reference,
  merchants info, change rates

- Source of truth for OLTP/OLAP/NoSQL

- Updates propagated through Change Data Captures (CDC)

- Holds sensitive information

  - field-level encryption
  - encrypted transfer (TLS 1.3)
  - strict access control and logging



merchants	
merchant_id 🔑	**bigint**
merchant_name	text
country_code	char(2)

customers	
customer_id 🔑	**bigint**
country_code	char(2)

currencies	
currency_code 🔑	**char(3)**
currency_name	text
usd_change_rate	float

payment_methods	
payment_method_id 🔑	**integer**
payment_method	text

payment_statuses	
payment_status_id 🔑	**integer**
payment_status	text

device_types	
device_type_id 🔑	**integer**
device_type	text

countries	
country_code 🔑	**char(2)**
currency_name	text

Online Transaction Processing (OLTP)

- Core transactional operations: payments processed through Stripe API

  High integrity, performance and reliabilty

- Normalized schema (3NF)

  Low redundancy, high consistency

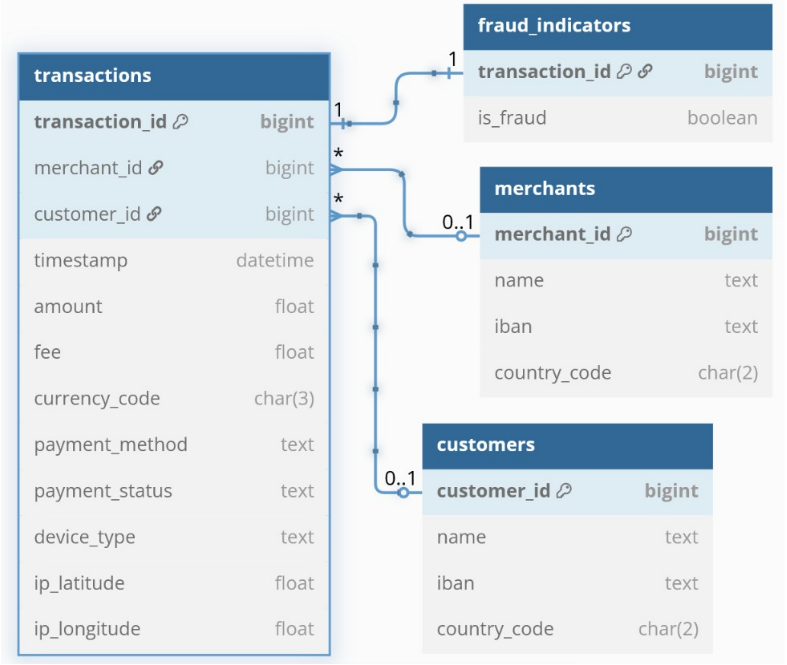
- ACID properties

  Atomicity, Consistency, Isolation, Durability

- Propagate downstream through Change Data Captures (CDC)

- Holds sensitive information

  - field-level encryption
  - encrypted transfer (TLS 1.3)
  - strict access control and logging



transactions	
transaction_id	**bigint**
merchant_id	bigint
customer_id	bigint
timestamp	datetime
amount	float
fee	float
currency_code	char(3)
payment_method	text
payment_status	text
device_type	text
ip_latitude	float
ip_longitude	float

fraud_indicators	
transaction_id	**bigint**
is_fraud	boolean

merchants	
merchant_id	**bigint**
name	text
iban	text
country_code	char(2)

customers	
customer_id	**bigint**
name	text
iban	text
country_code	char(2)

Online Analytical Processing (OLAP)

- Analytical queries and BI

  High availability, fast queries

- Star schema architecture

  `transactions` fact table

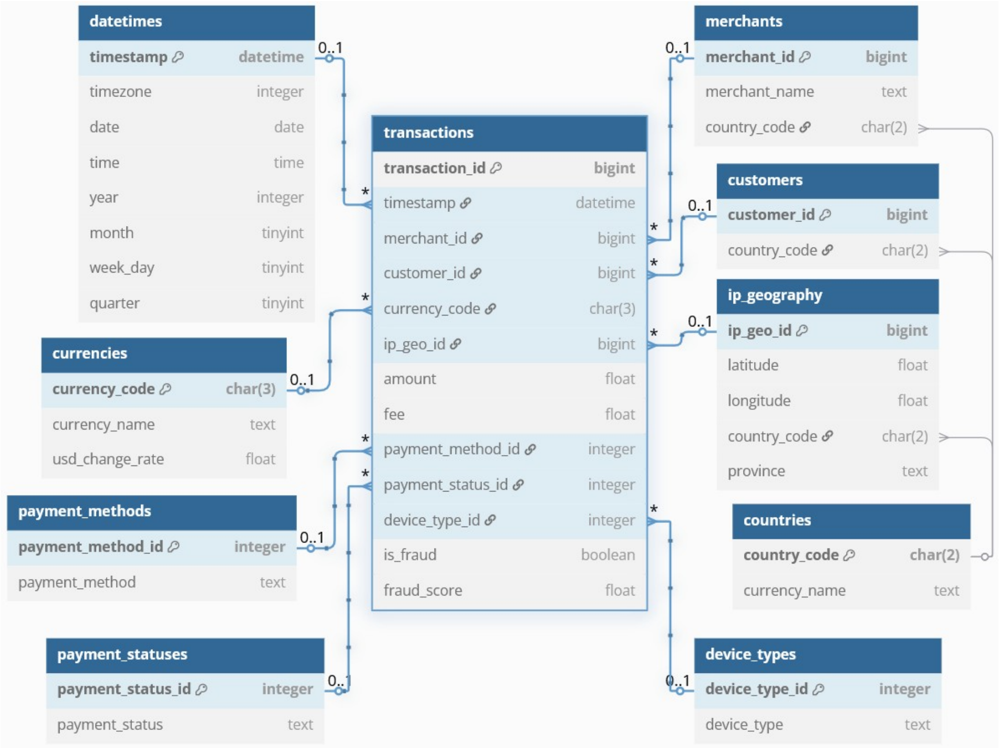
- Pre-aggregations, views

  for performance optimization

- Connected to analytics and machine learning components

- Less-sensitive information

  - Anonymized data
  - Access control and logging



datetimes	
timestamp	**datetime**
timezone	integer
date	date
time	time
year	integer
month	tinyint
week_day	tinyint
quarter	tinyint

transactions	
transaction_id	**bigint**
timestamp	datetime
merchant_id	bigint
customer_id	bigint
currency_code	char(3)
ip_geo_id	bigint
amount	float
fee	float
payment_method_id	integer
payment_status_id	integer
device_type_id	integer
is_fraud	boolean
fraud_score	float

merchants	
merchant_id	**bigint**
merchant_name	text
country_code	char(2)

customers	
customer_id	**bigint**
country_code	char(2)

ip_geography	
ip_geo_id	**bigint**
latitude	float
longitude	float
country_code	char(2)
province	text

countries	
country_code	**char(2)**
currency_name	text

currencies	
currency_code	**char(3)**
currency_name	text
usd_change_rate	float

payment_methods	
payment_method_id	**integer**
payment_method	text

payment_statuses	
payment_status_id	**integer**
payment_status	text

device_types	
device_type_id	**integer**
device_type	text

NoSQL Database

- Document-based for semi-structure and unstructured data

  Telemetry, customer feedback, logs, ML features

- Embedded documents

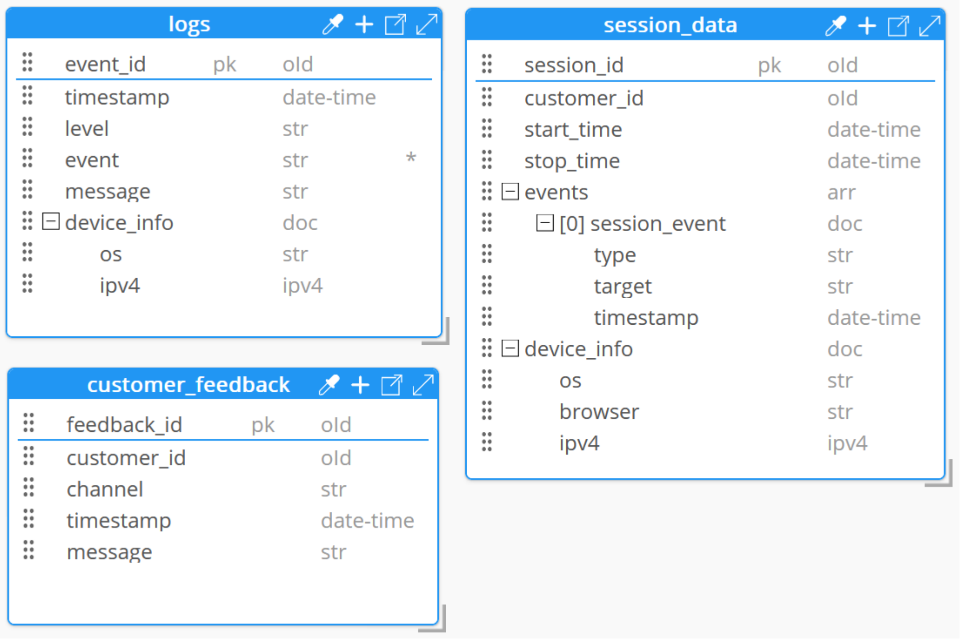
  for tightly coupled data (e.g. device info)

- Document referencing

  to link large documents
  (e.g. session data and customer feedback)

- No sensitive information stored

  - Anonymized data
  - Access control and logging



logs

event_id	pk	oId	
timestamp		date-time	
level		str	
event		str	*
message		str	
⊟ device_info		doc	
os		str	
ipv4		ipv4	

customer_feedback

feedback_id	pk	oId
customer_id		oId
channel		str
timestamp		date-time
message		str

session_data

session_id	pk	oId
customer_id		oId
start_time		date-time
stop_time		date-time
⊟ events		arr
⊟ [0] session_event		doc
type		str
target		str
timestamp		date-time
⊟ device_info		doc
os		str
browser		str
ipv4		ipv4

Compliance

- Compliance with international regulations (PCI-DSS, GDPR, etc)
- Confidential information in Reference and OLTP databases

  Encryption, strict access policy and logging
- No sensitive data in OLAP and NoSQL databases

  Anonymous or tokenized data
- Creation of encrypted backups

Compliance

- Compliance with international regulations (PCI-DSS, GDPR, etc)
- Confidential information in Reference and OLTP databases

  Encryption, strict access policy and logging
- No sensitive data in OLAP and NoSQL databases

  Anonymous or tokenized data
- Creation of encrypted backups

ML Integration

- Features extraction from OLAP and NoSQL databases

  Batch orchestration with Apache Airflow
- Model management with MLflow

  Traceability, version control
- Online inference through APIs

  e.g. for fraud scoring, queried by OLAP
- Deployment within Kubernetes

  High-availability infrastructure
- Monitoring with Evidently

  Data drift, performance degradation

Thanks!