Designing Observability for Fault Diagnosis

Elton Stoneman
Consultant & Trainer

@EltonStoneman   blog.sixeyed.com

GLOBOTICKET

A Globomantics Company

Amila

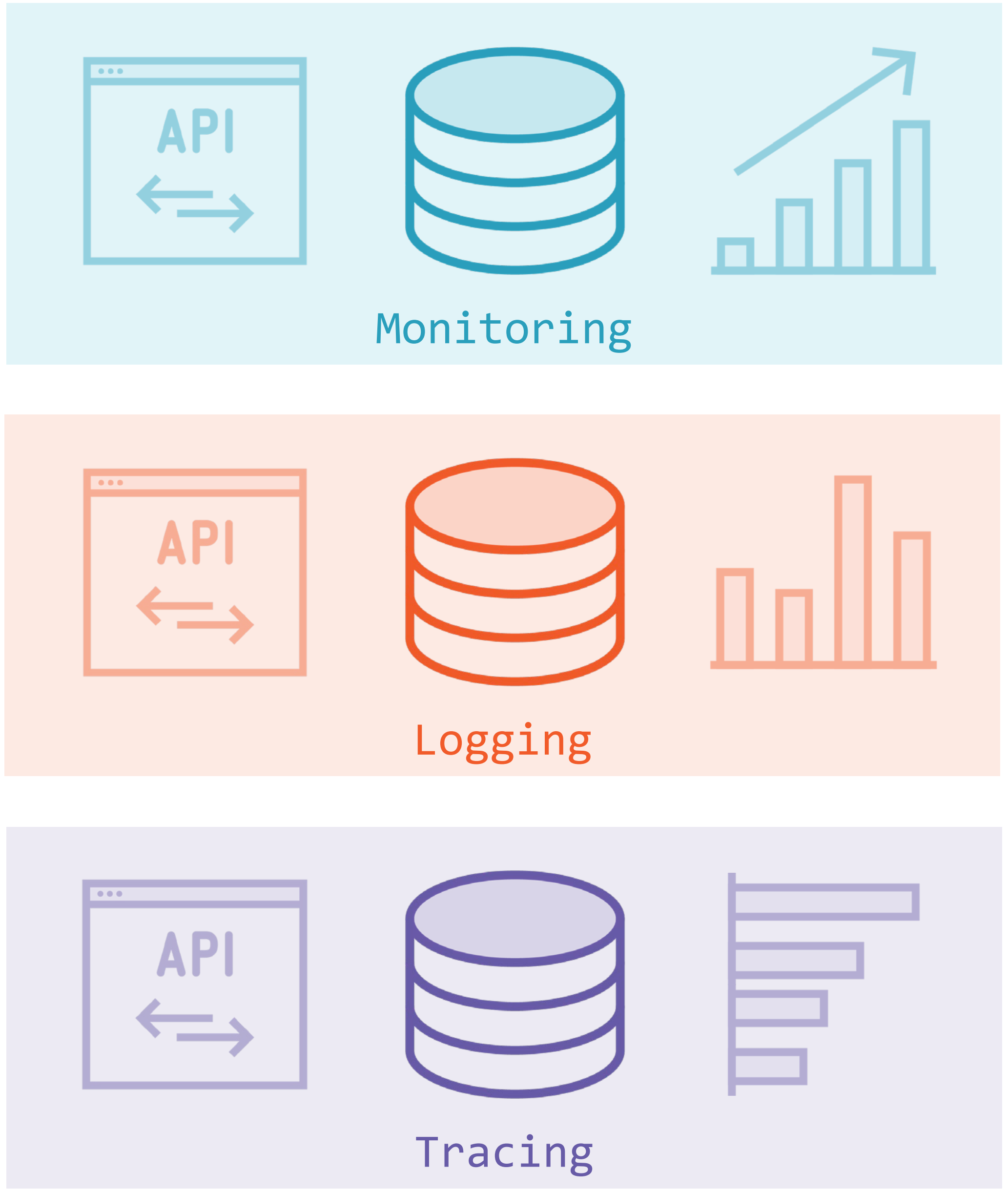
API
Monitoring

API
Logging

API
Tracing

GLOBOTICKET

A Globomantics Company



Incident Model

Triage **Examine** Diagnose Test Cure

Triage

- Get back to "good enough"
- Application health dashboards
- High-level metrics overview

Examine

- Understand the problem & trigger
- Application logs & network traces
- Low-level details

Golden Signals

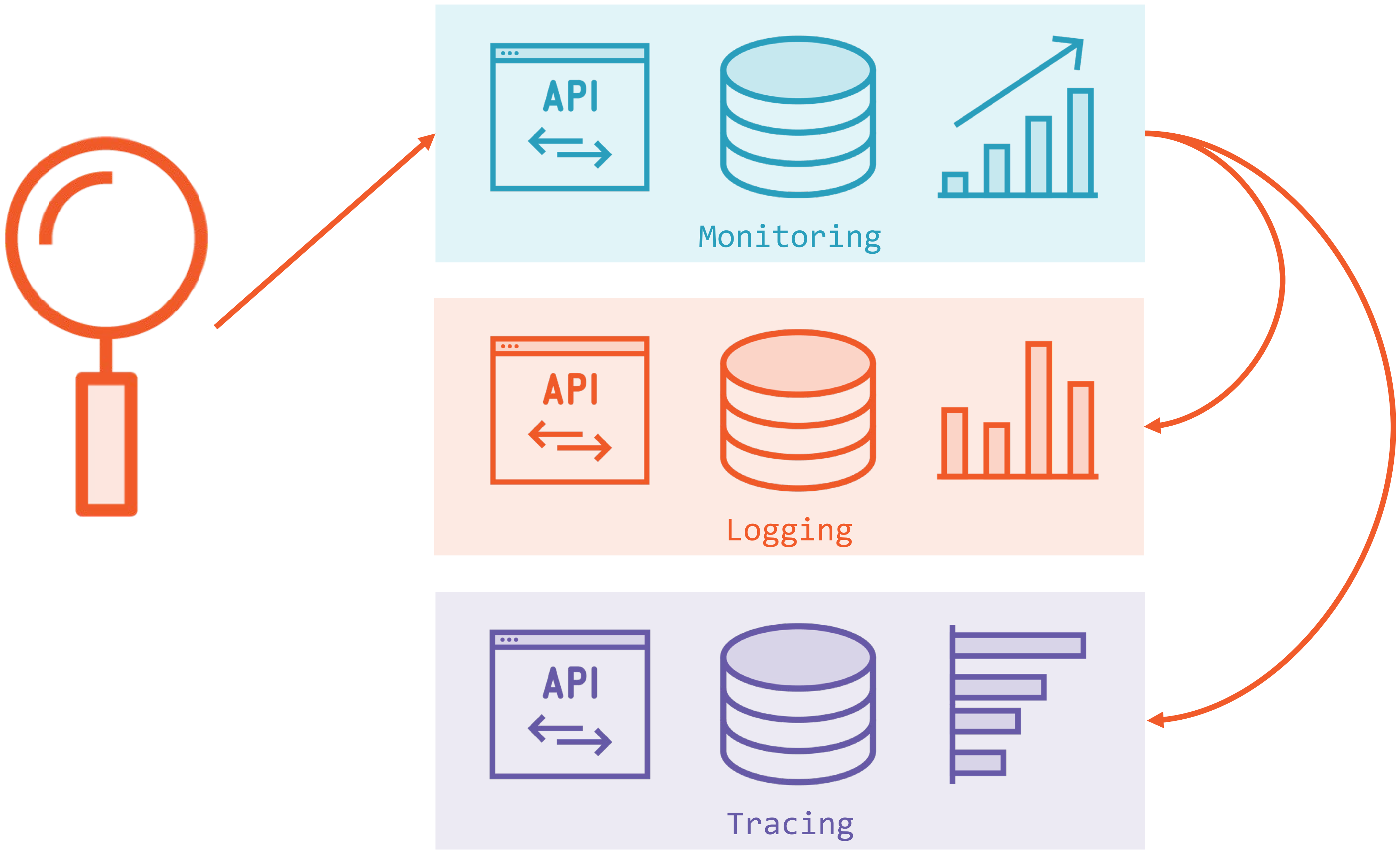
Latency	Job processing time Response generation time
Traffic	Length of message queue Requests per second
Errors	Request failures Response correctness
Saturation	CPU & memory utilization Network bandwidth

Observability Data

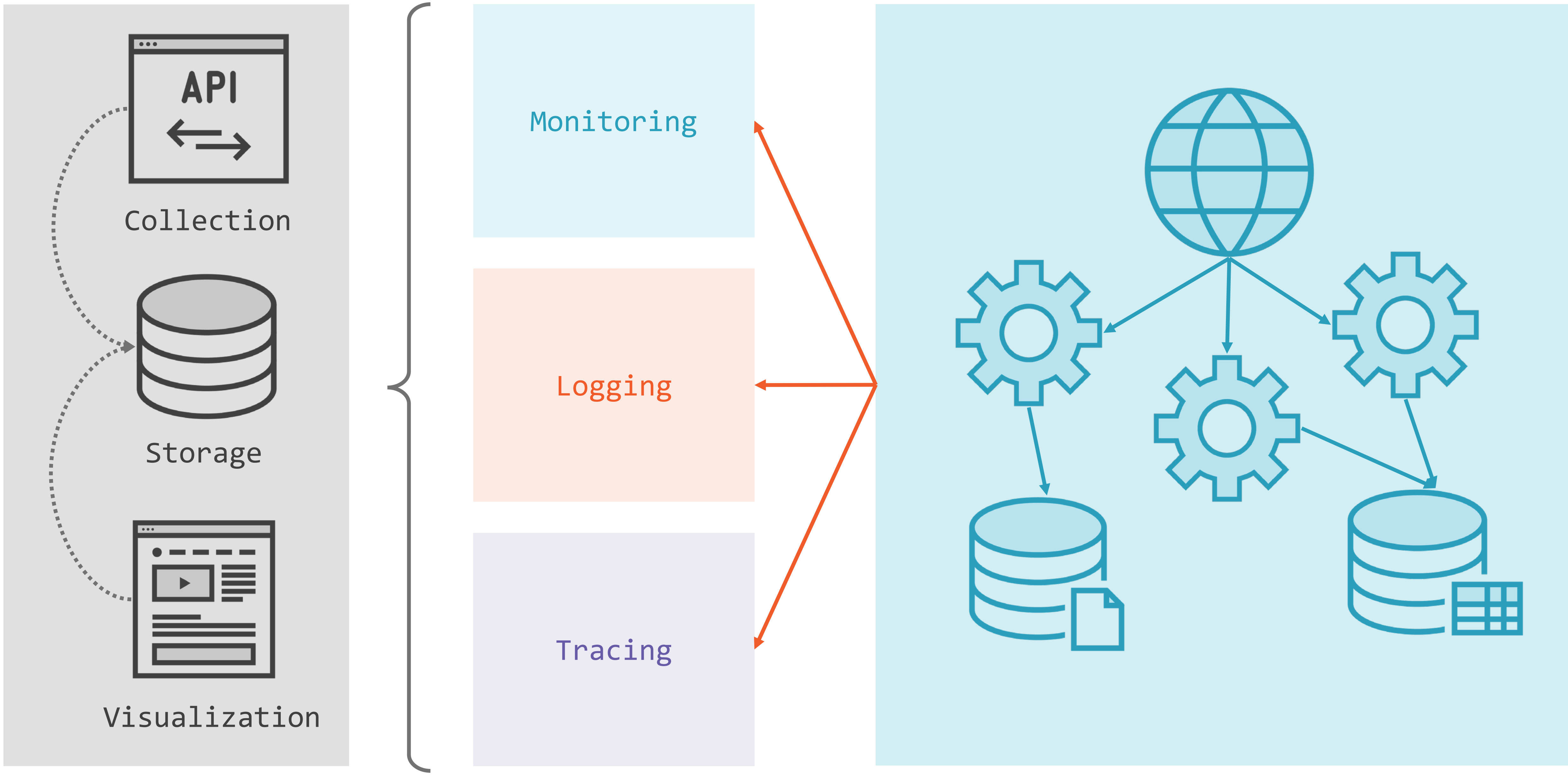
Detail levels

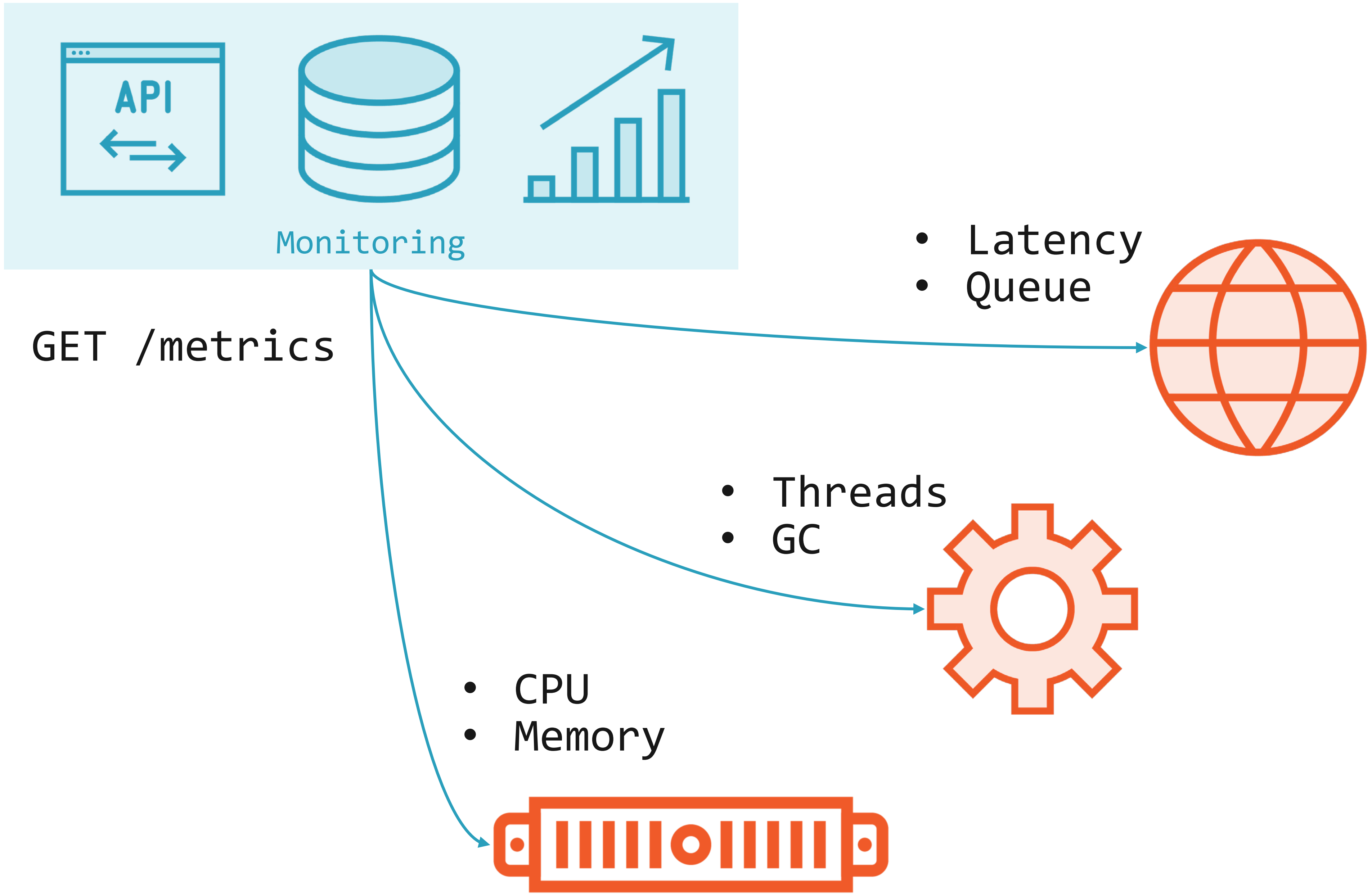
Amounts available

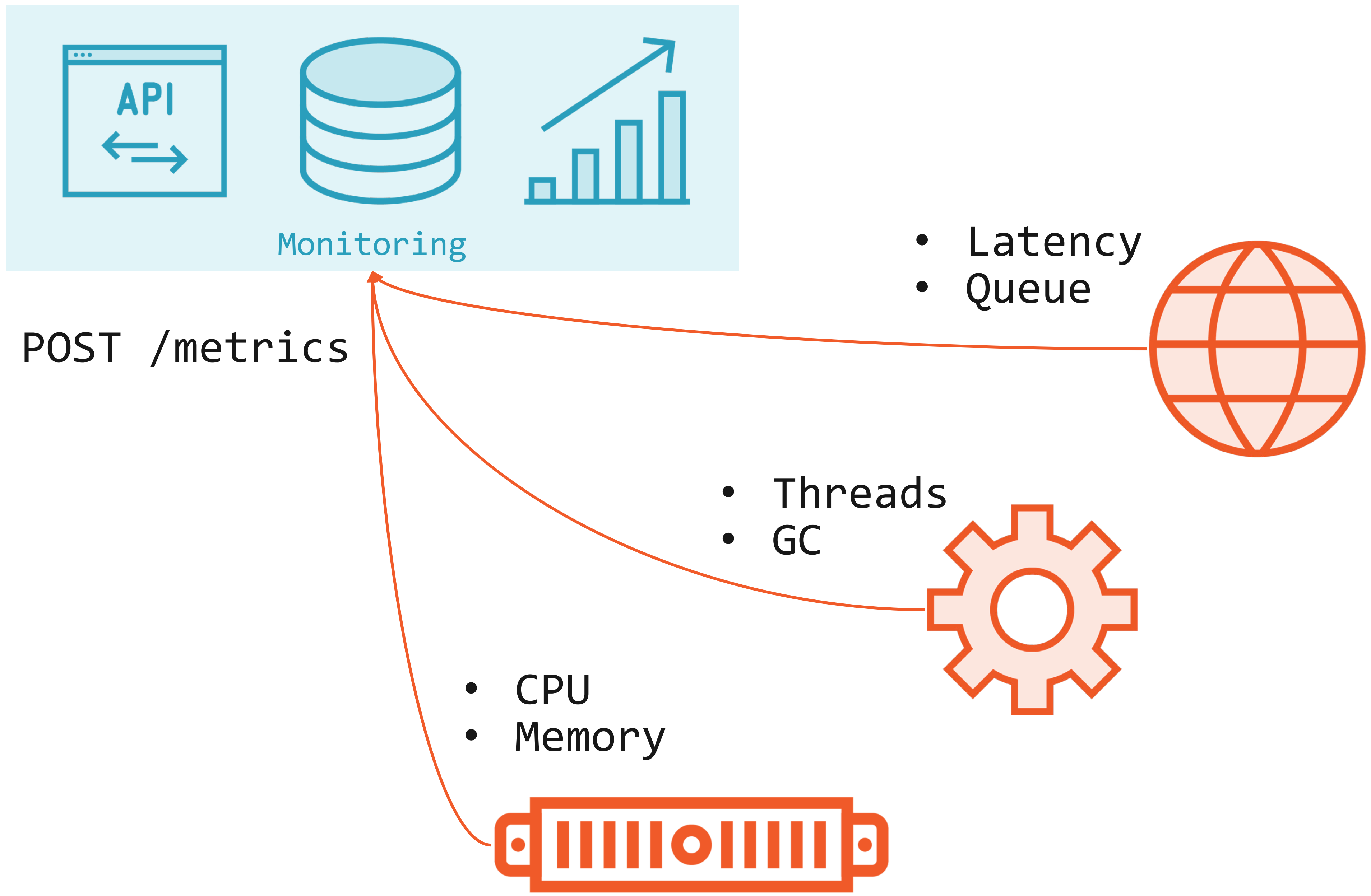
Investigative tools

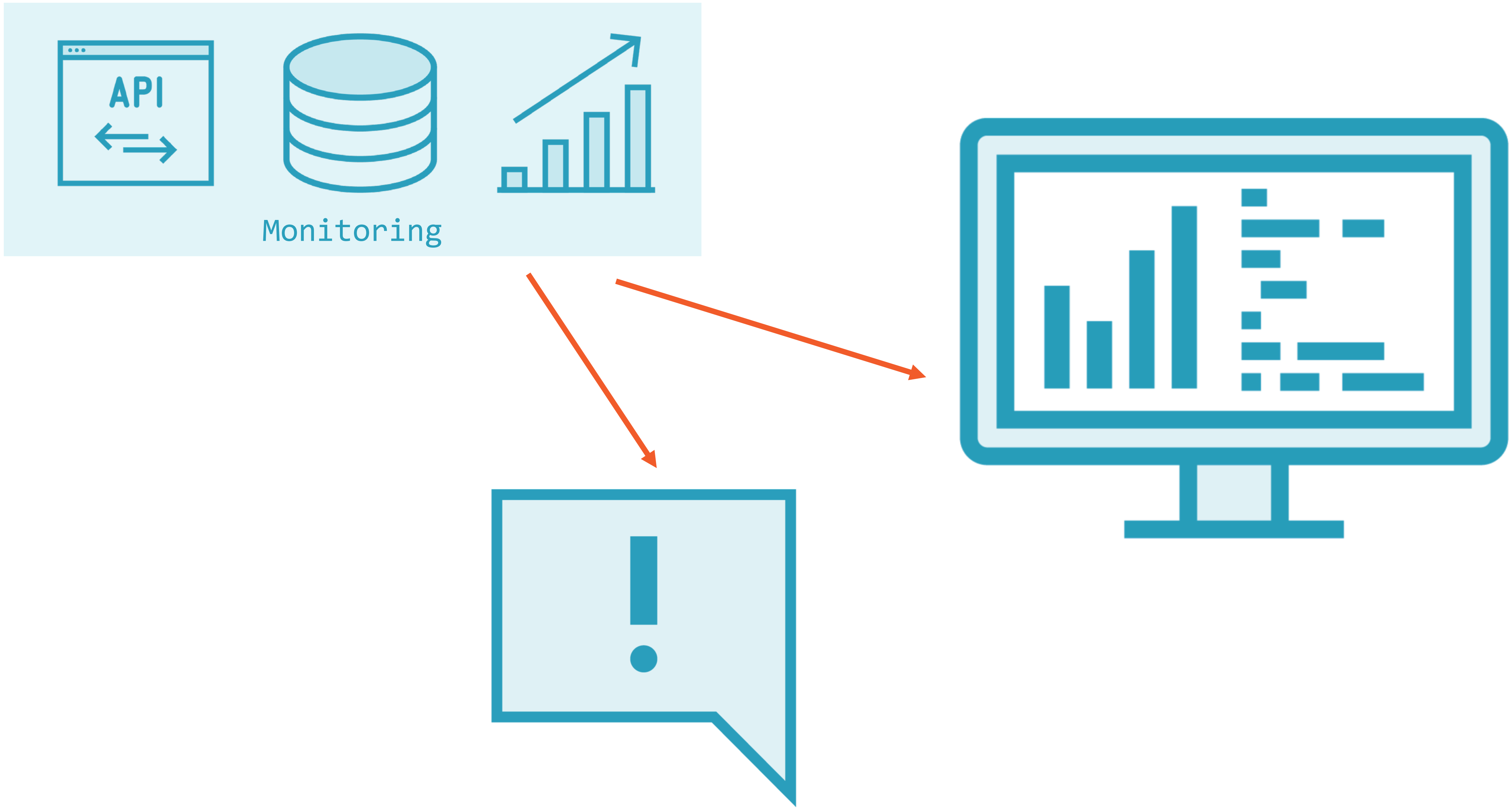

API
Monitoring
API
Logging
API
Tracing

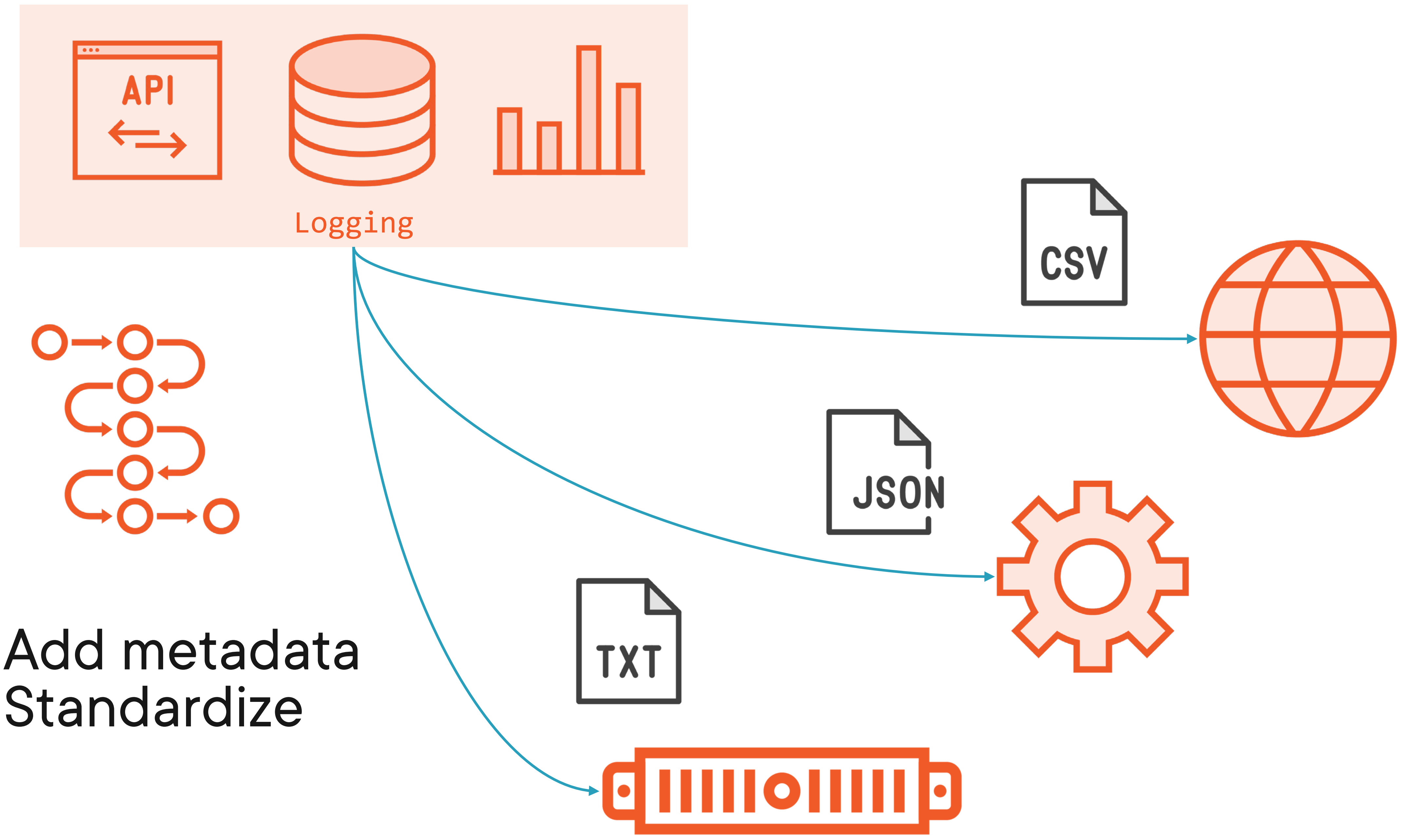
Exploring the Three Pillars of Observability


API
Collection
Storage
Visualization
Monitoring
Logging
Tracing

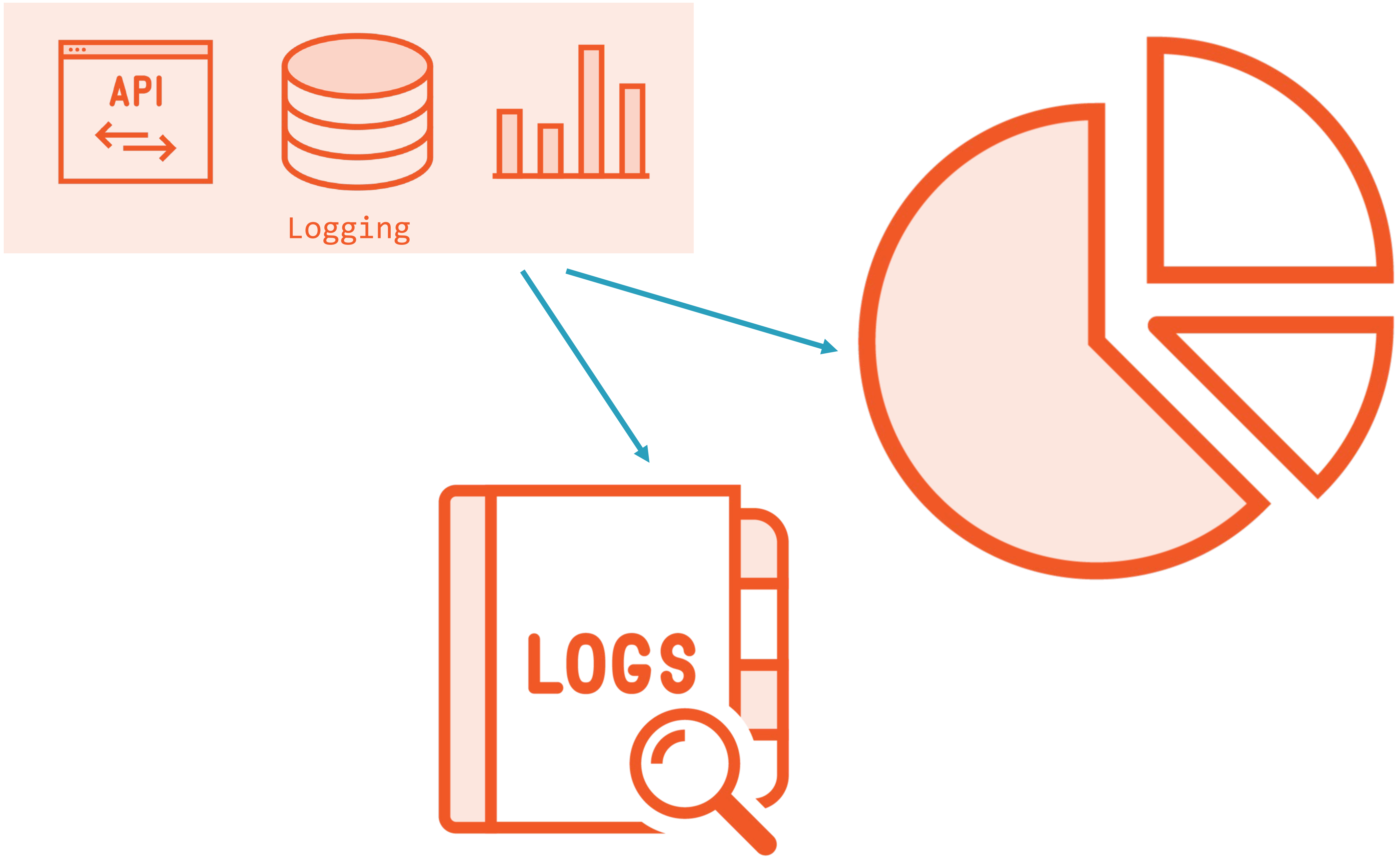

API
Monitoring
GET /metrics
• Latency
• Queue
• Threads
• GC
• CPU
• Memory

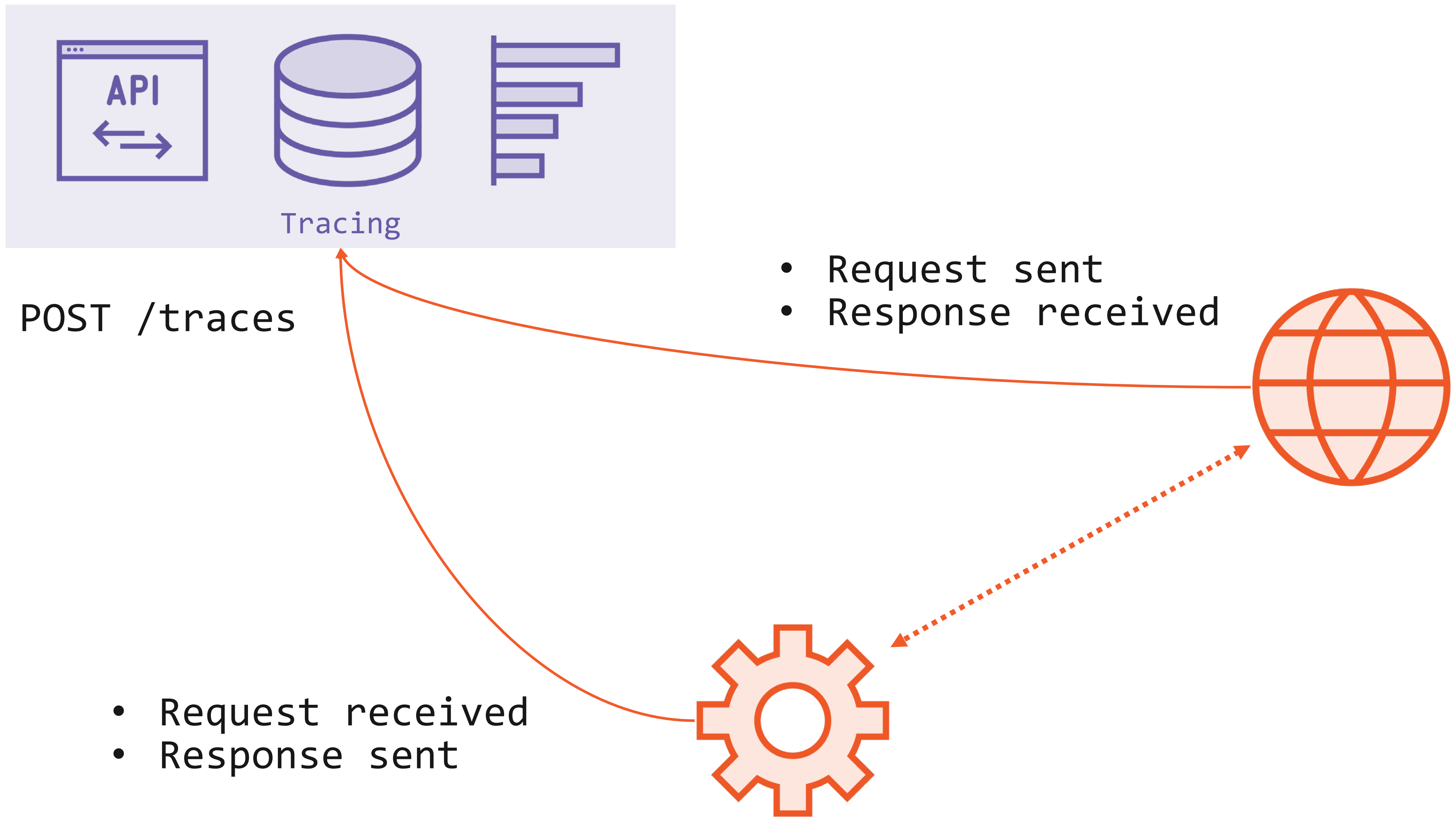

API
Monitoring
POST /metrics
• Latency
• Queue
• Threads
• GC
• CPU
• Memory

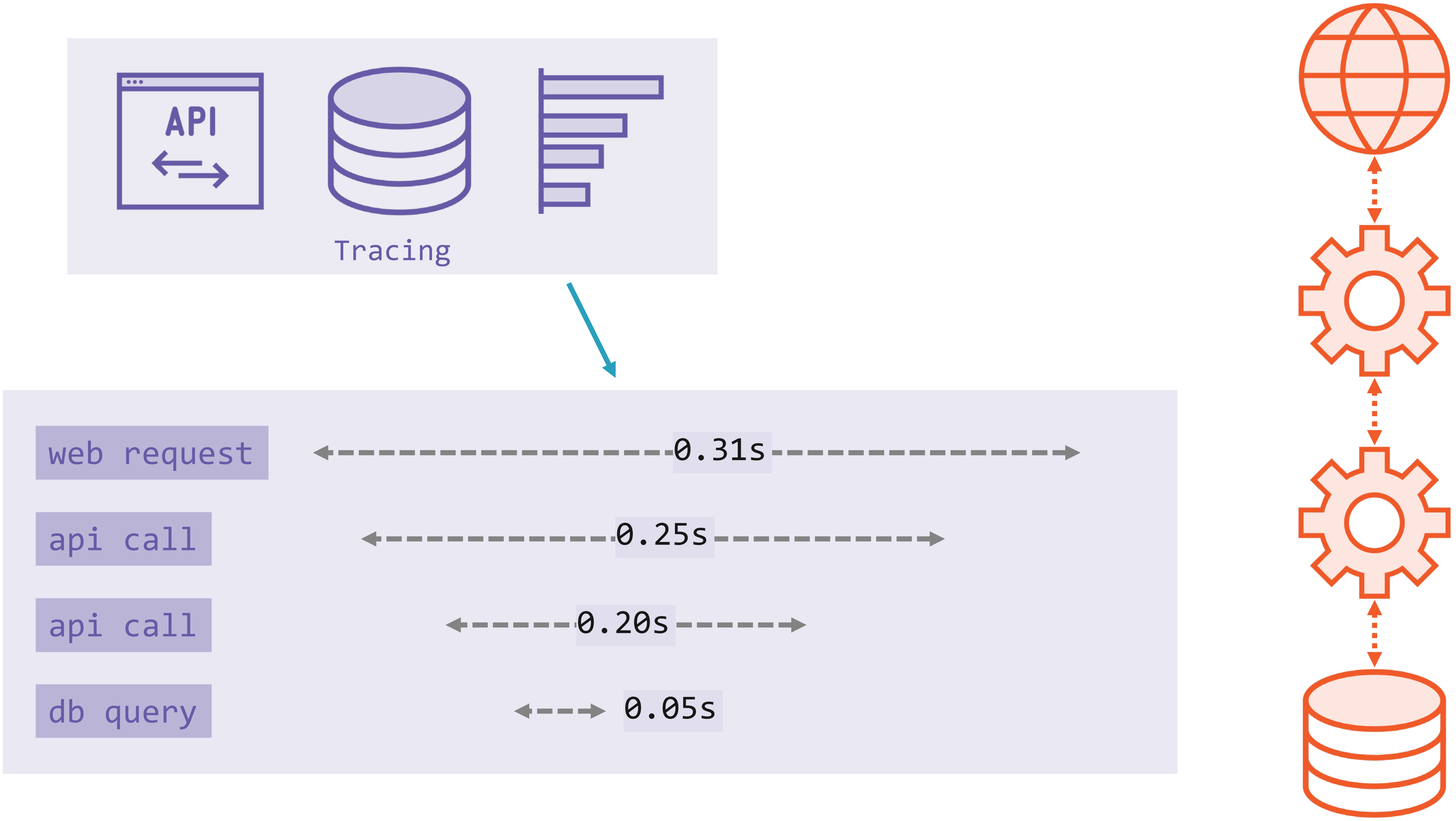

API
Monitoring



- Add metadata
- Standardize


API
Logging
LOGS


API
Tracing
POST /traces
• Request sent
• Response received
• Request received
• Response sent


API
Tracing
web request
0.31s
api call
0.25s
api call
0.20s
db query
0.05s

#1

- Overall insight
- Uptime & usage
- App-level metrics

#2

- Low-level detail
- Collect existing logs
- Transform pipeline

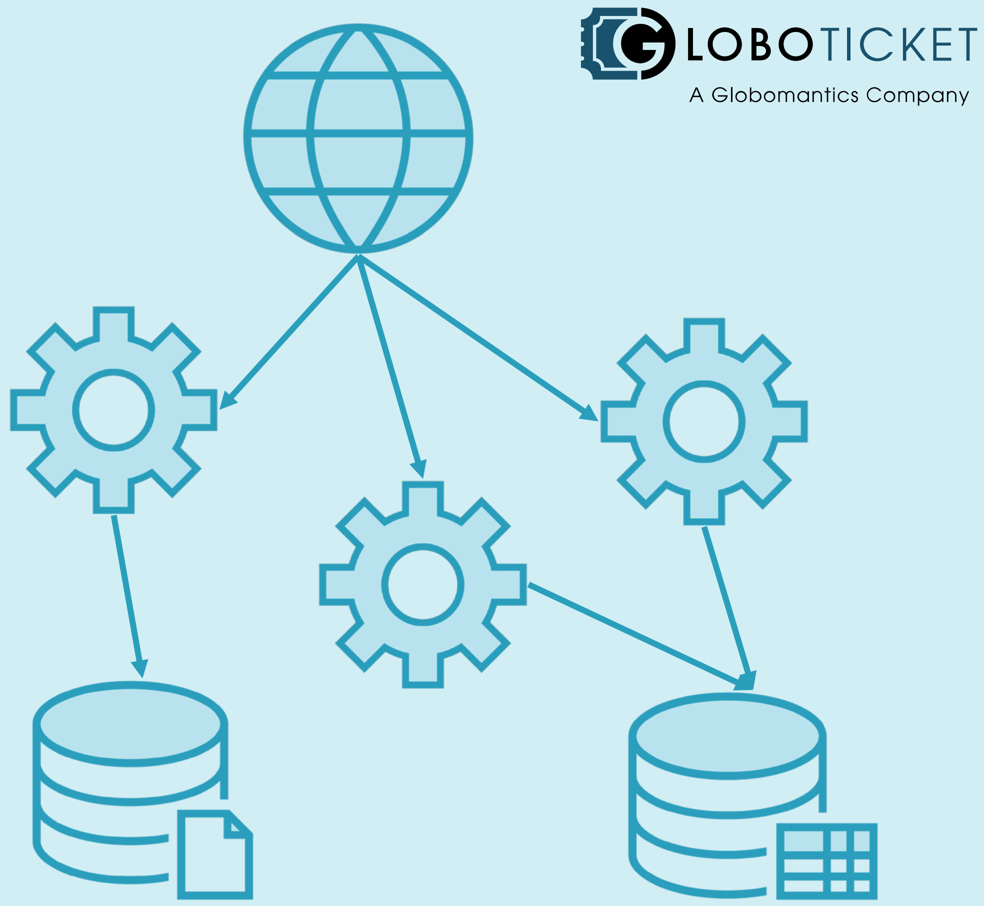
#3

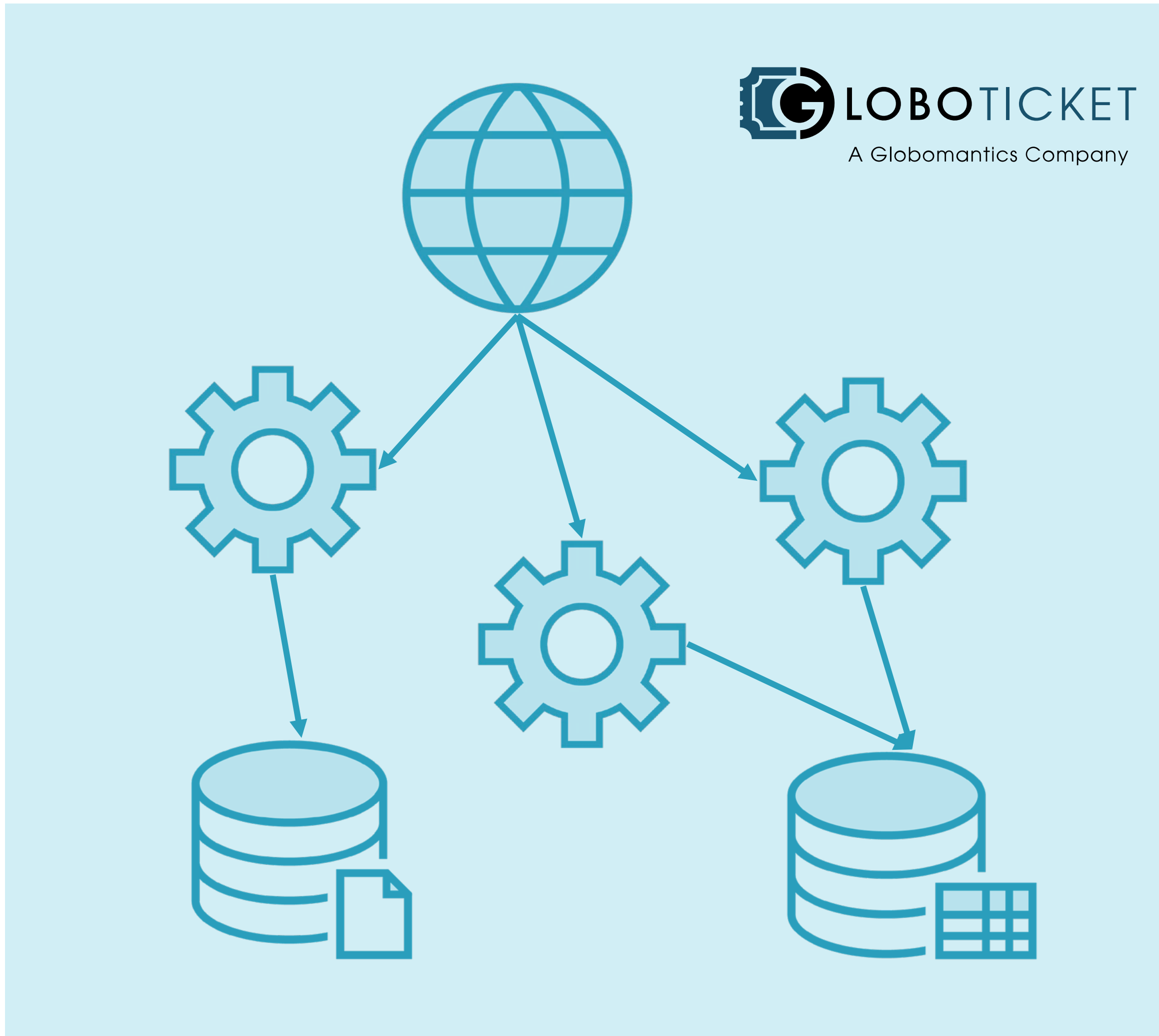
- Communication detail
- Visualize interaction
- Needs tracing library

Scenario: Putting Metrics to Use

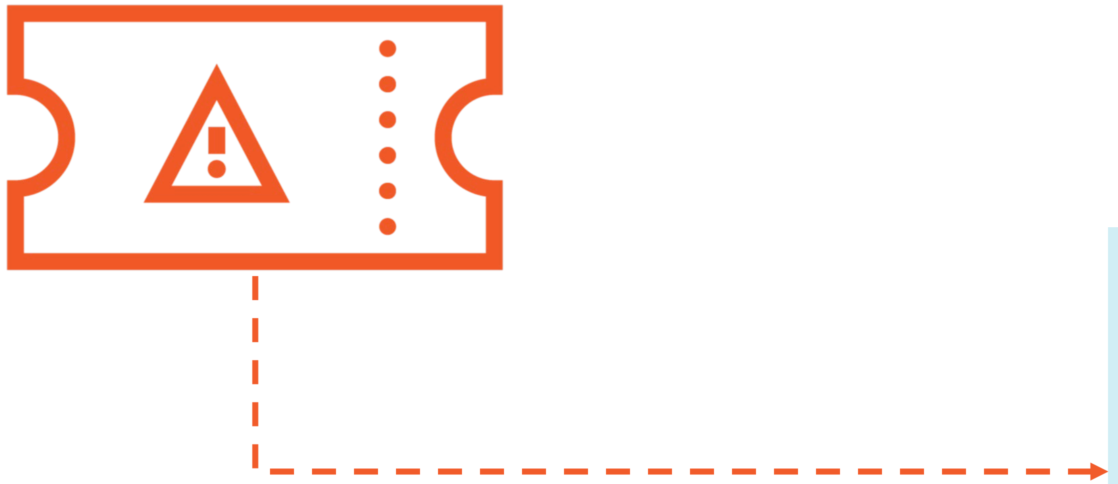
Amila

How do you add capacity when you need to?

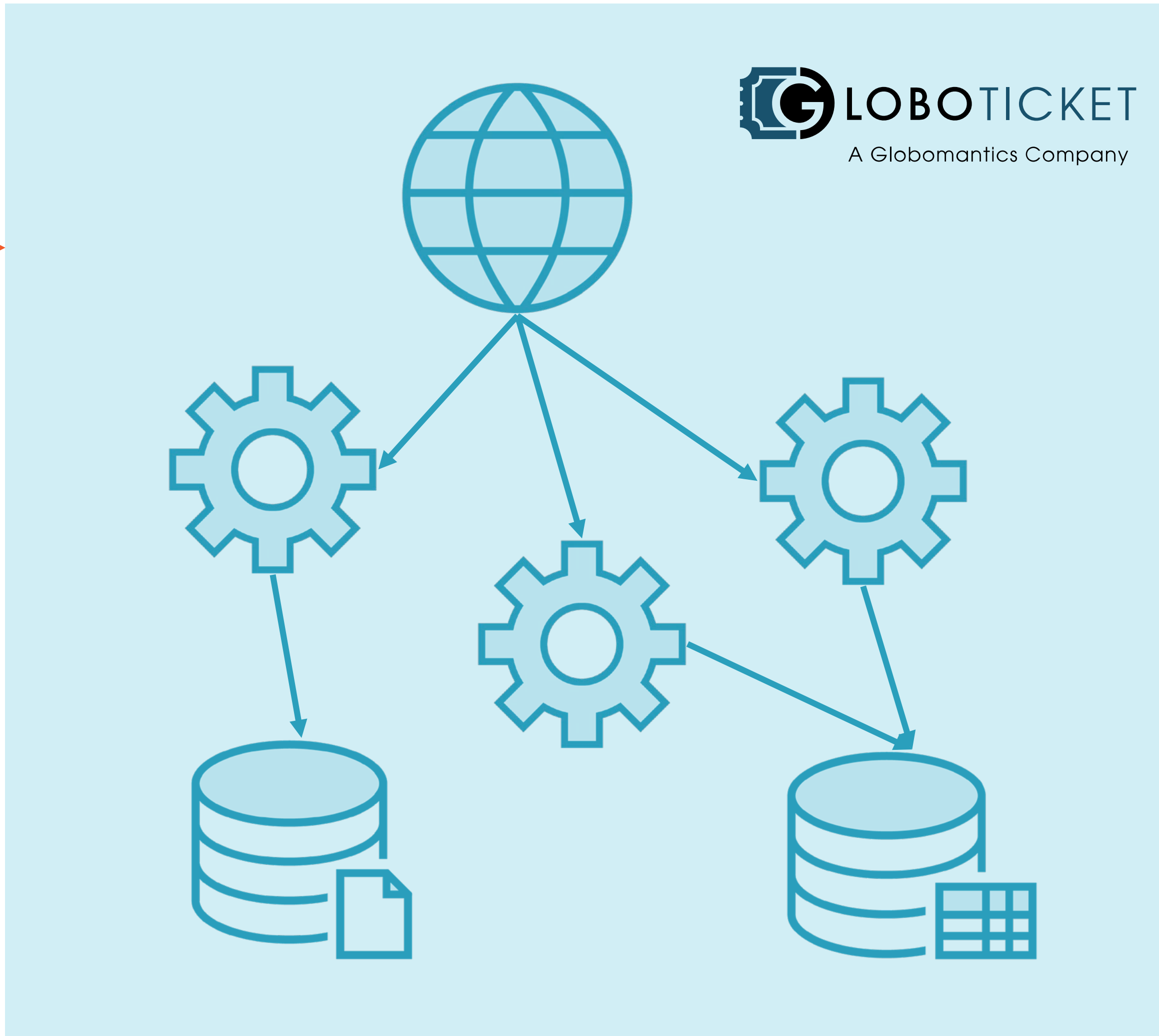


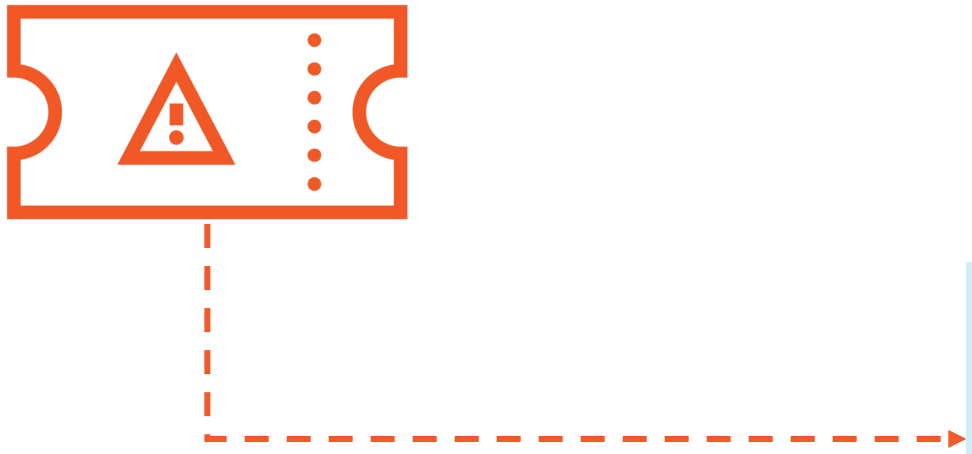


- CPU & memory measured
- Tracked by component
- Alerts outside normal range

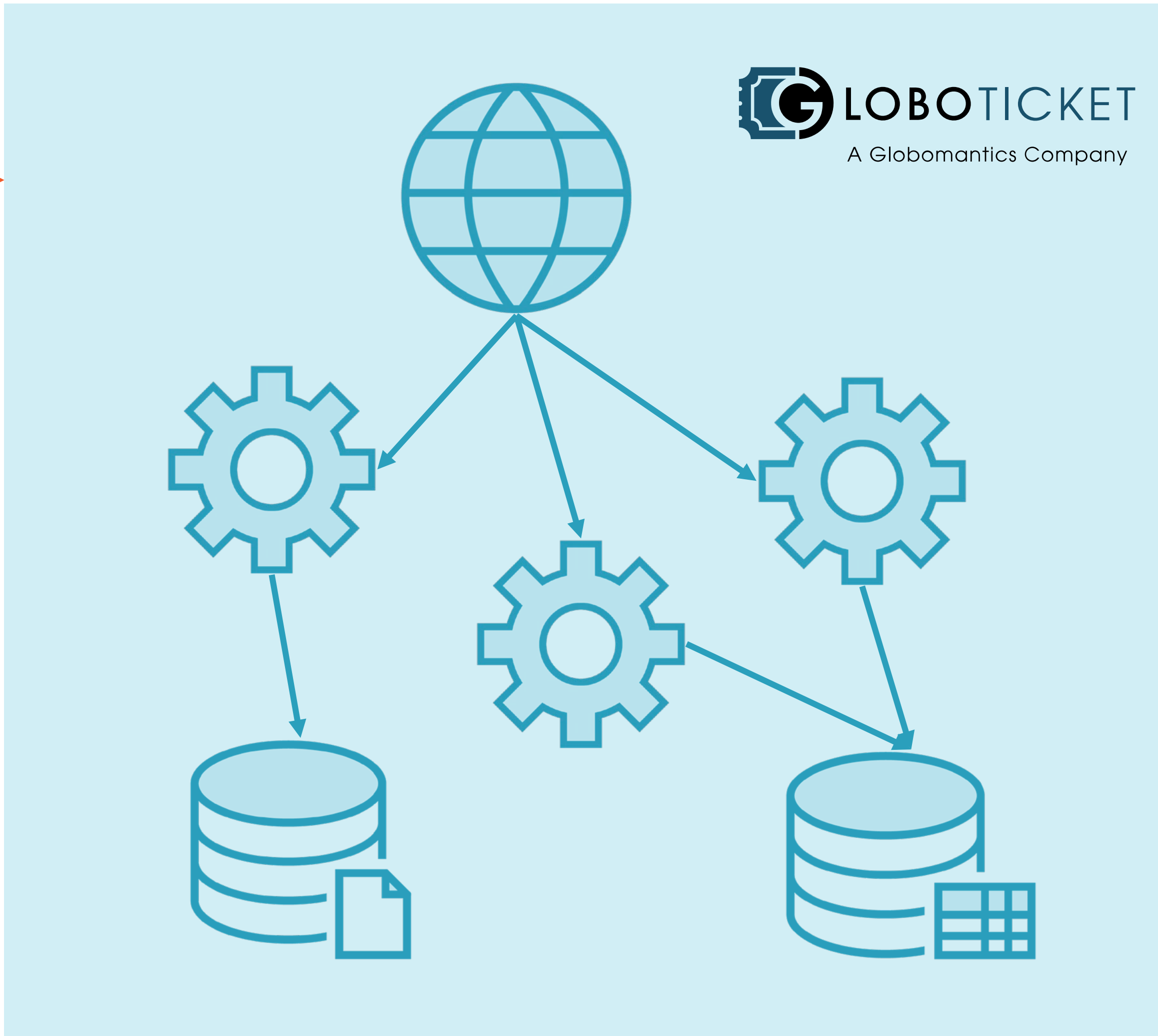


- Already auto-scaled
- Engineer investigates
- Adds more scale if needed

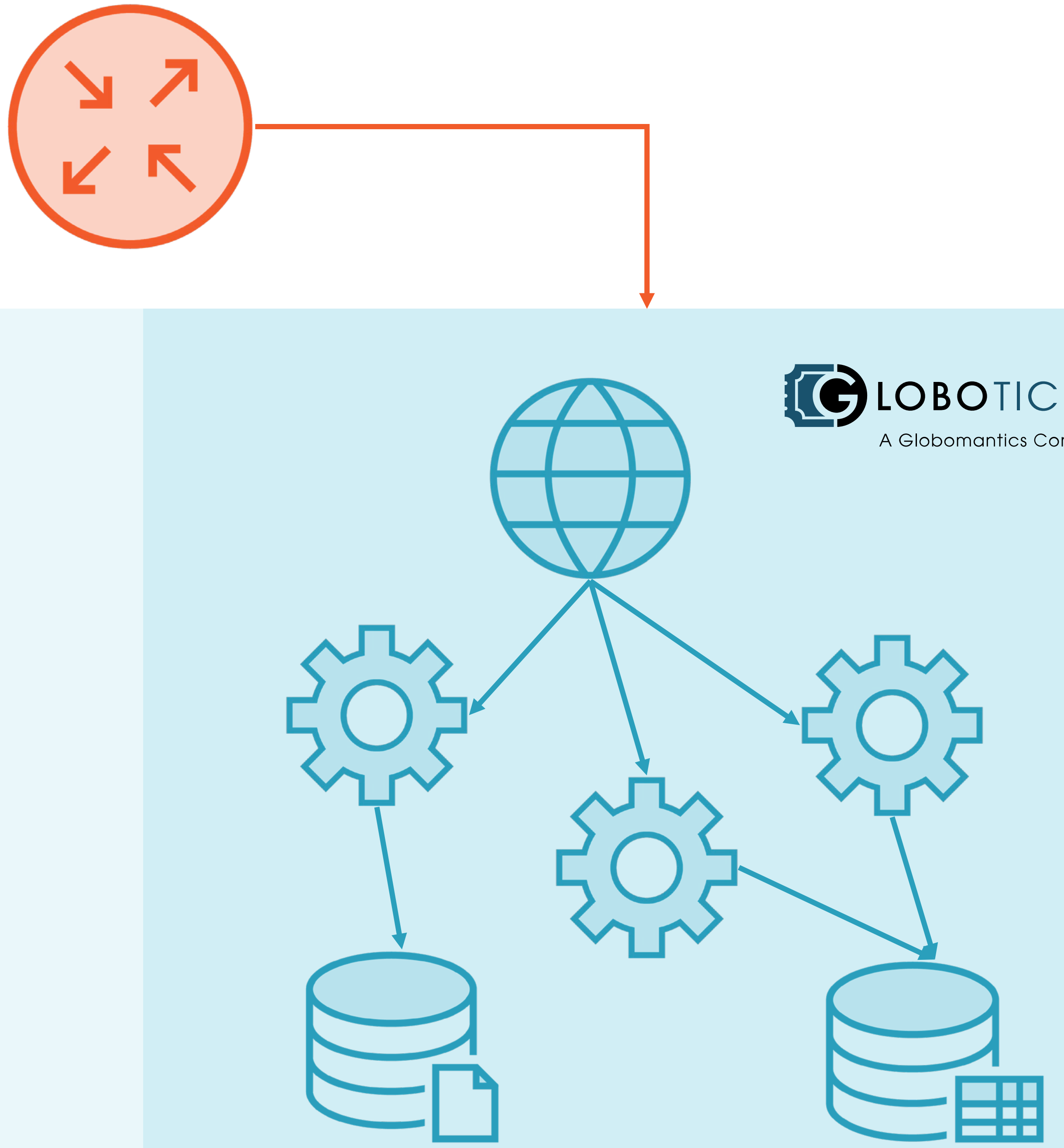




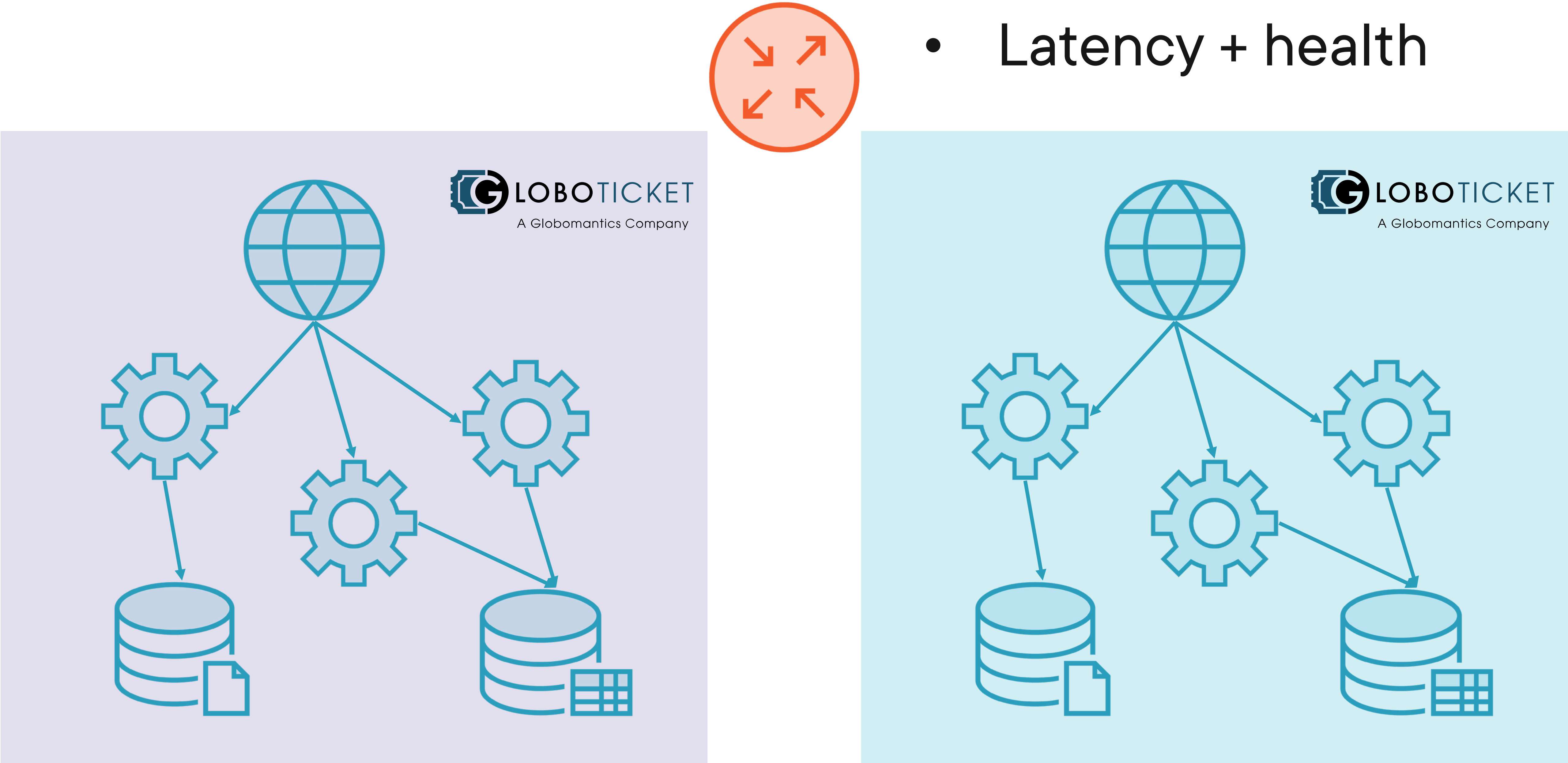
- Alert on budget
- Standard runbook
- Scale guidance



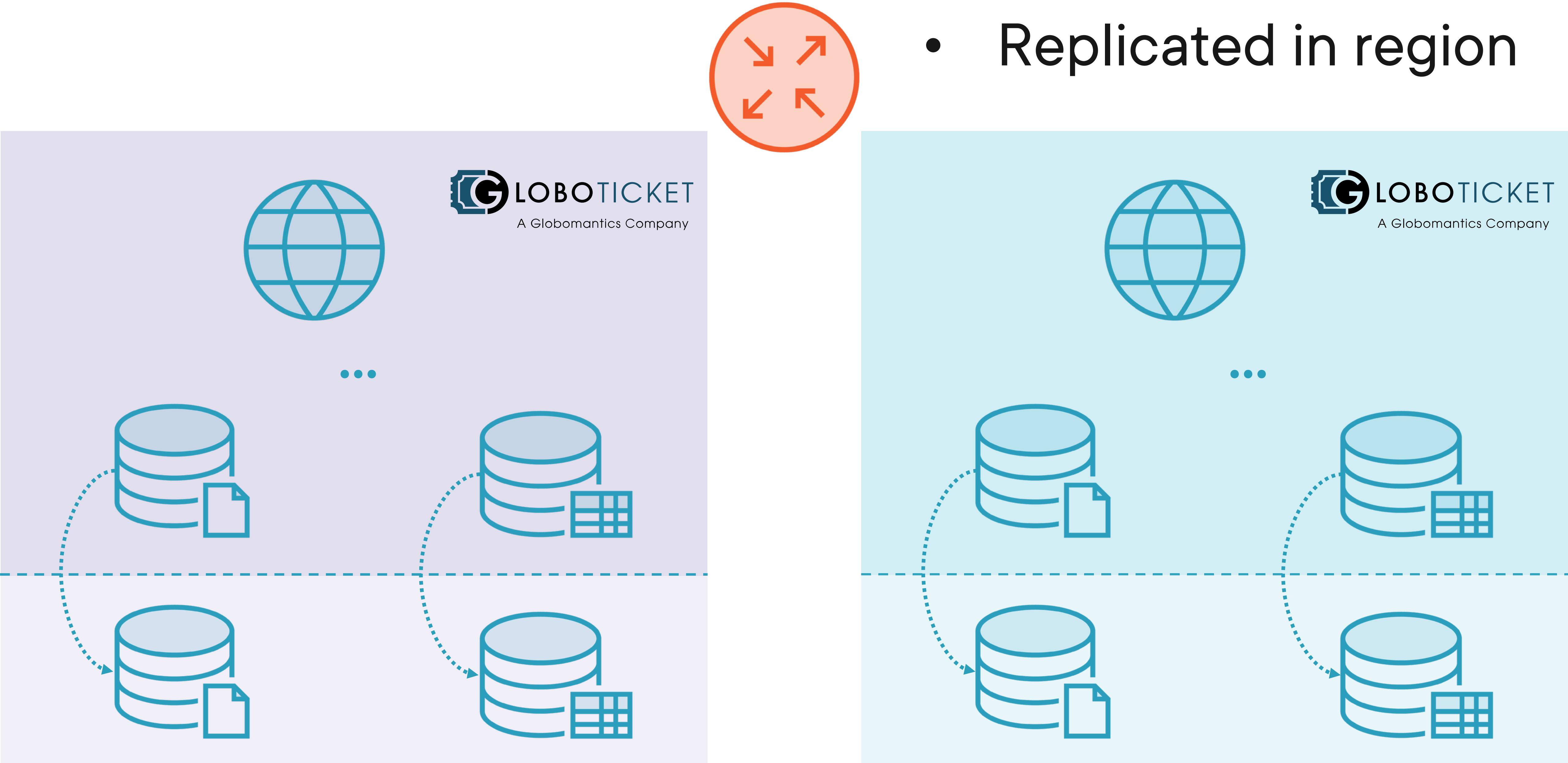
Can you shift traffic between regions?



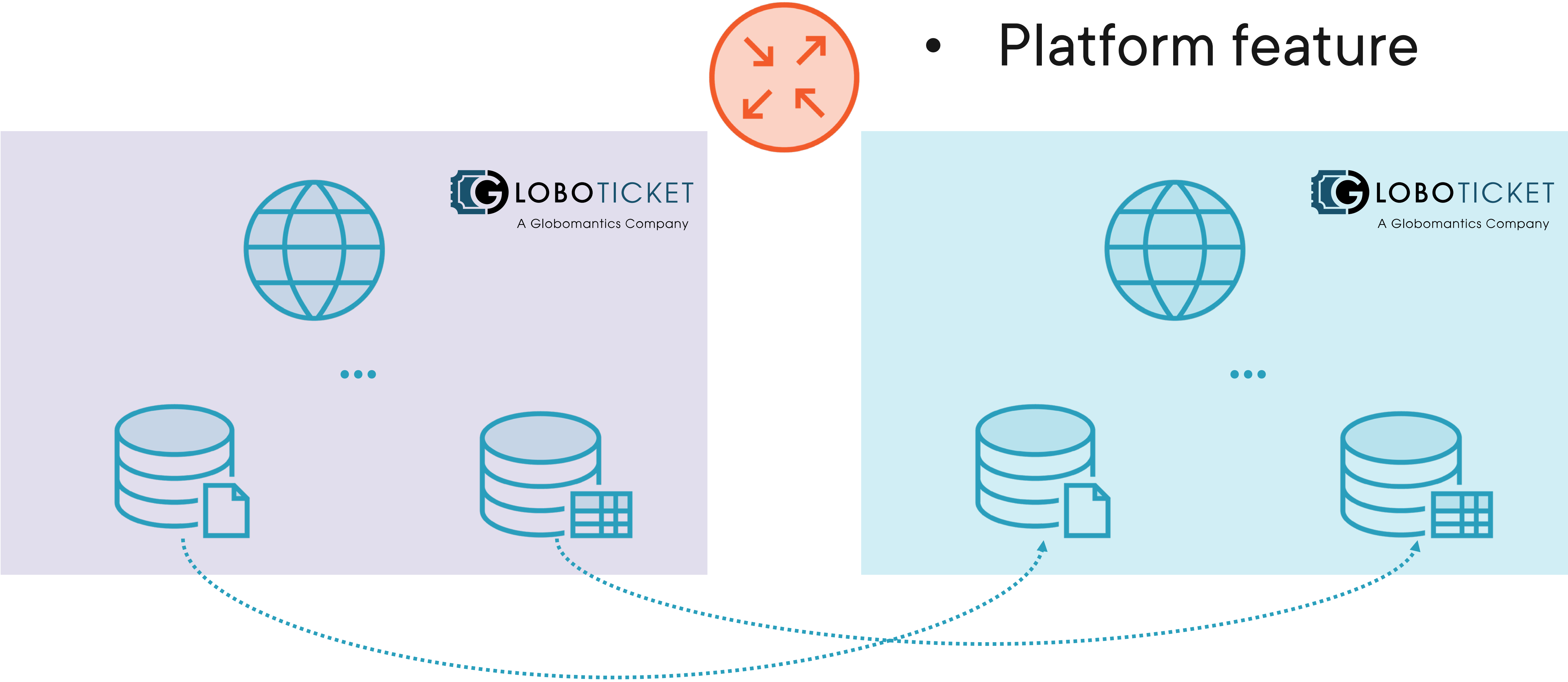
- DNS routing
- Latency + health



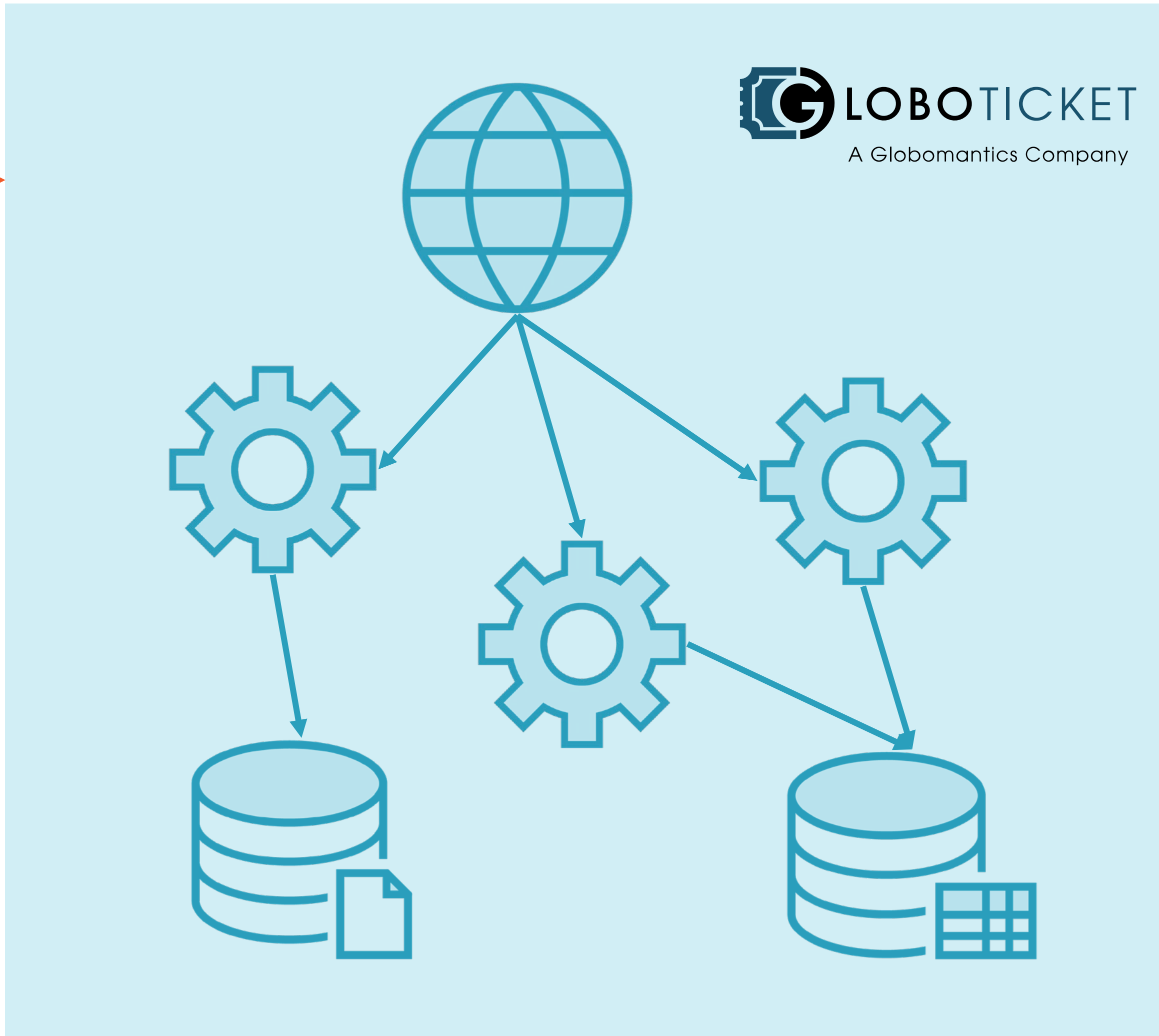
- Region-specific data
- Replicated in region



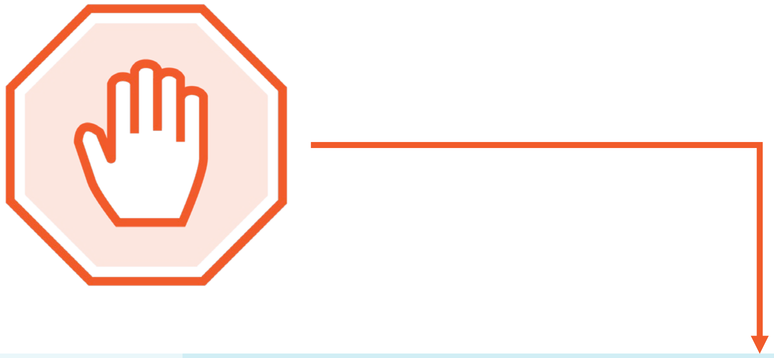
- Global replication
- Platform feature

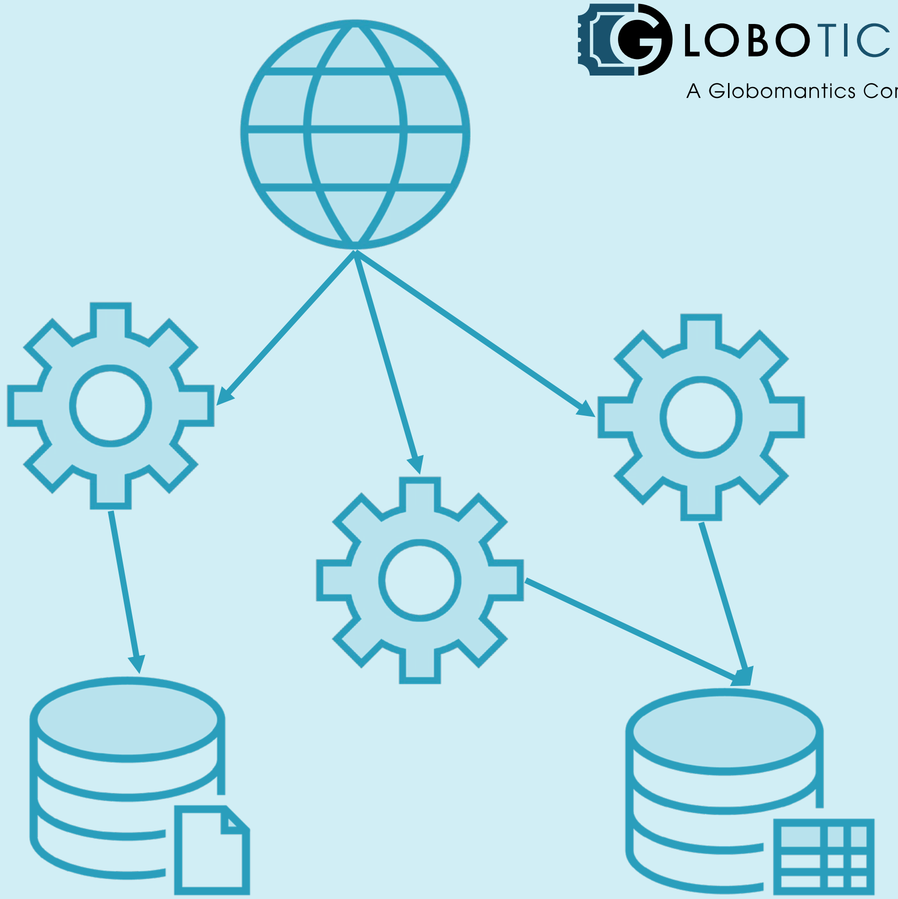


- Global data
- Routing metrics
- Manual override

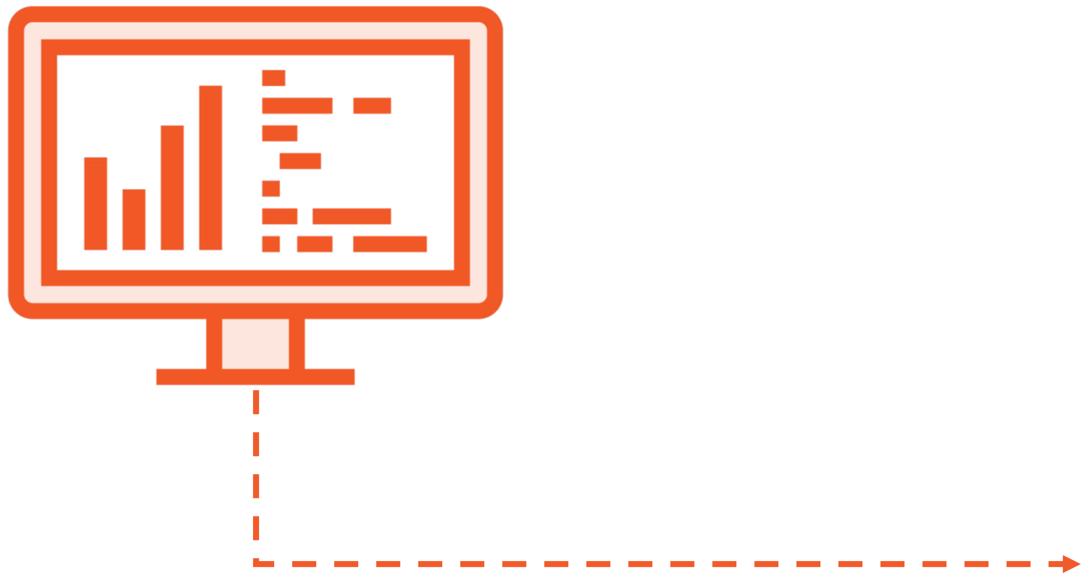


When do you downgrade service levels?

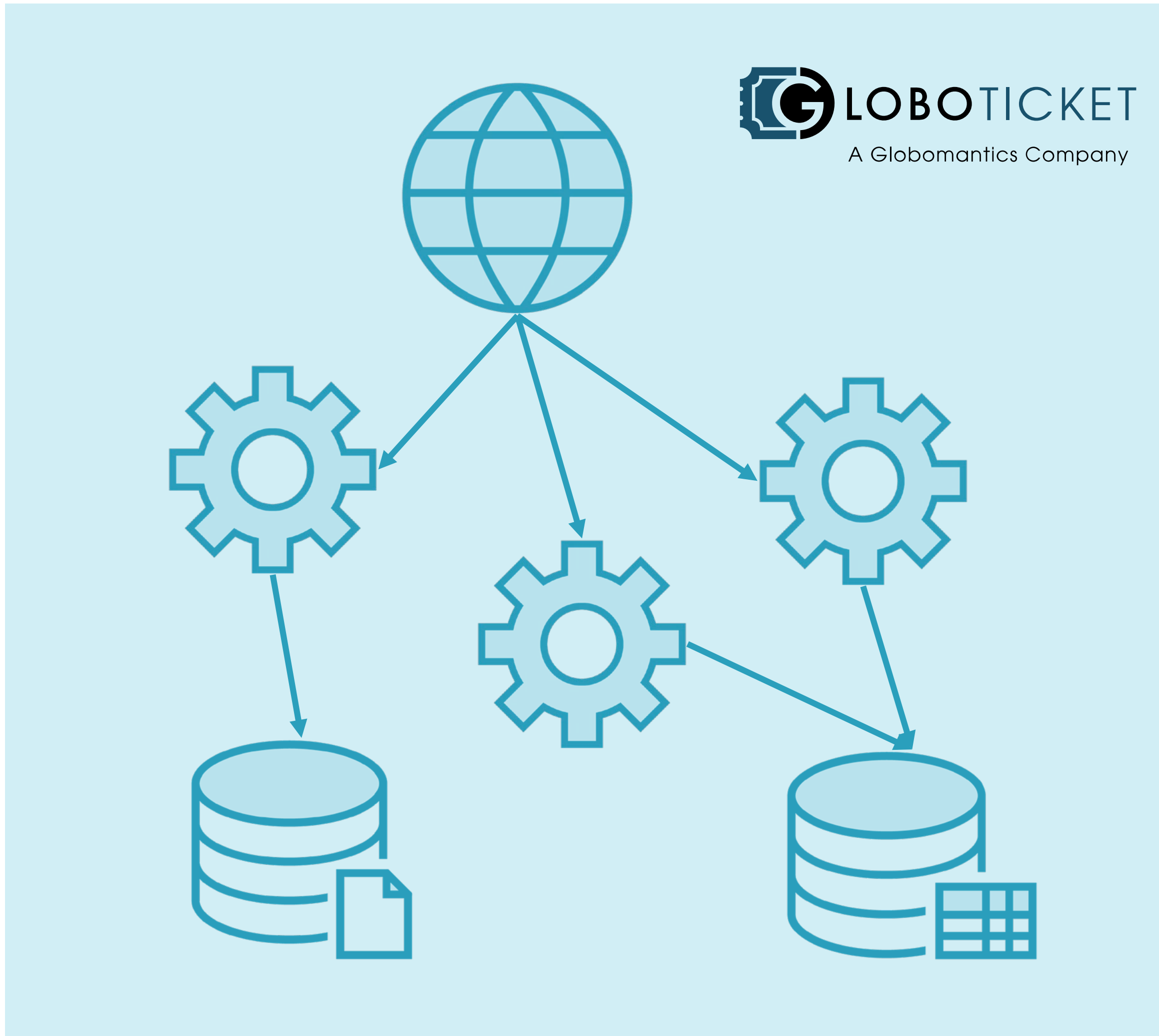


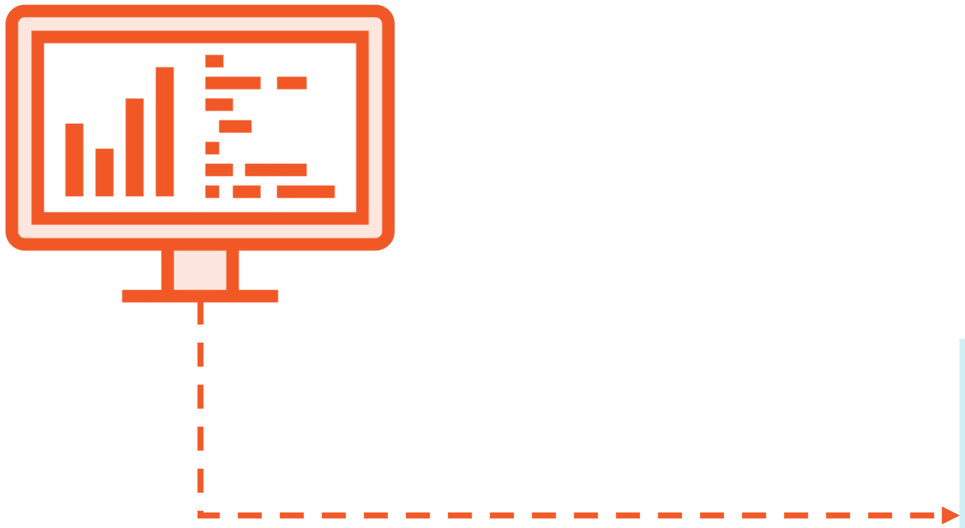


LOBOTICKET
A Globomantics Company

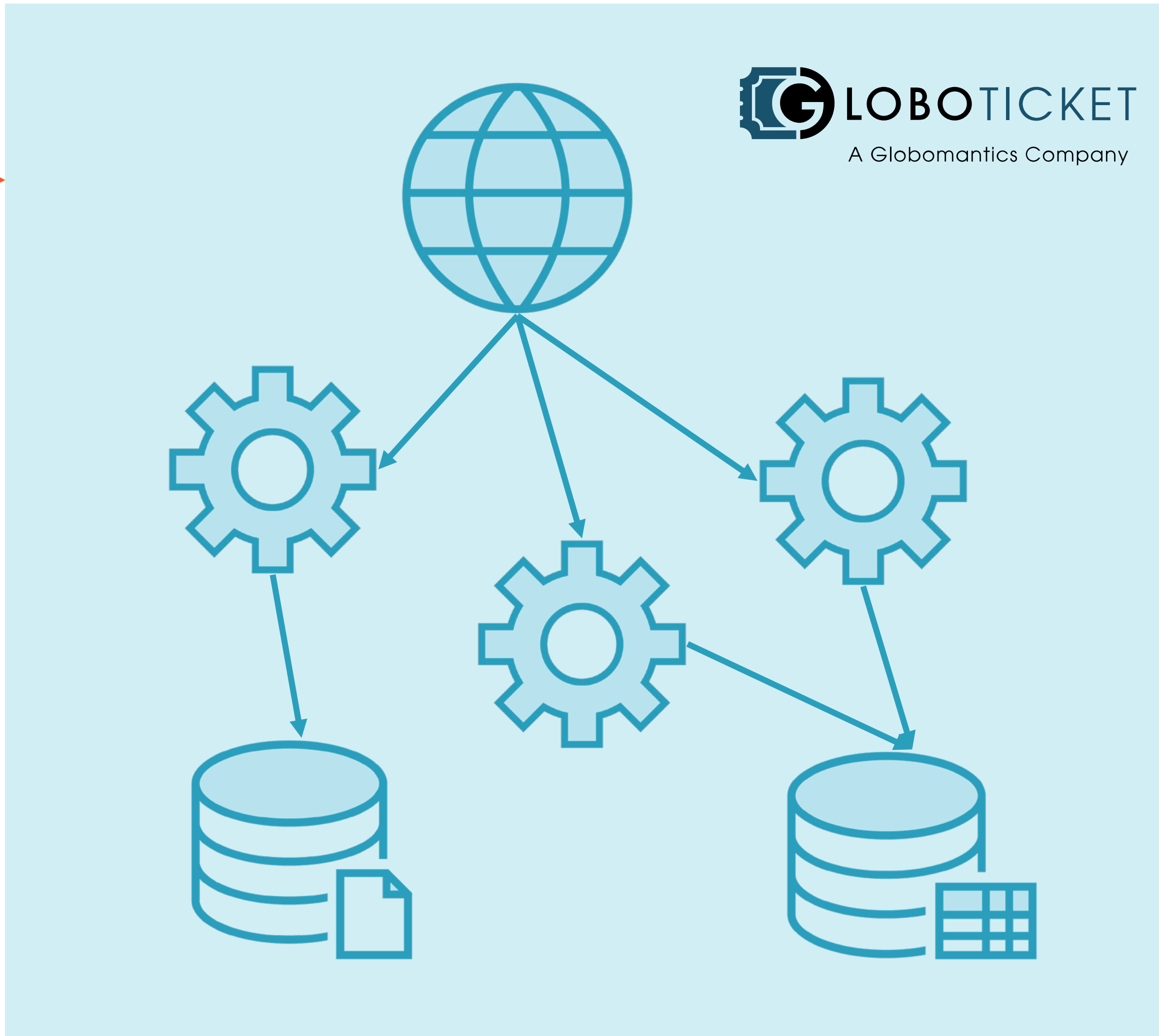


- Component saturation?
- Region saturation?
- Global saturation?

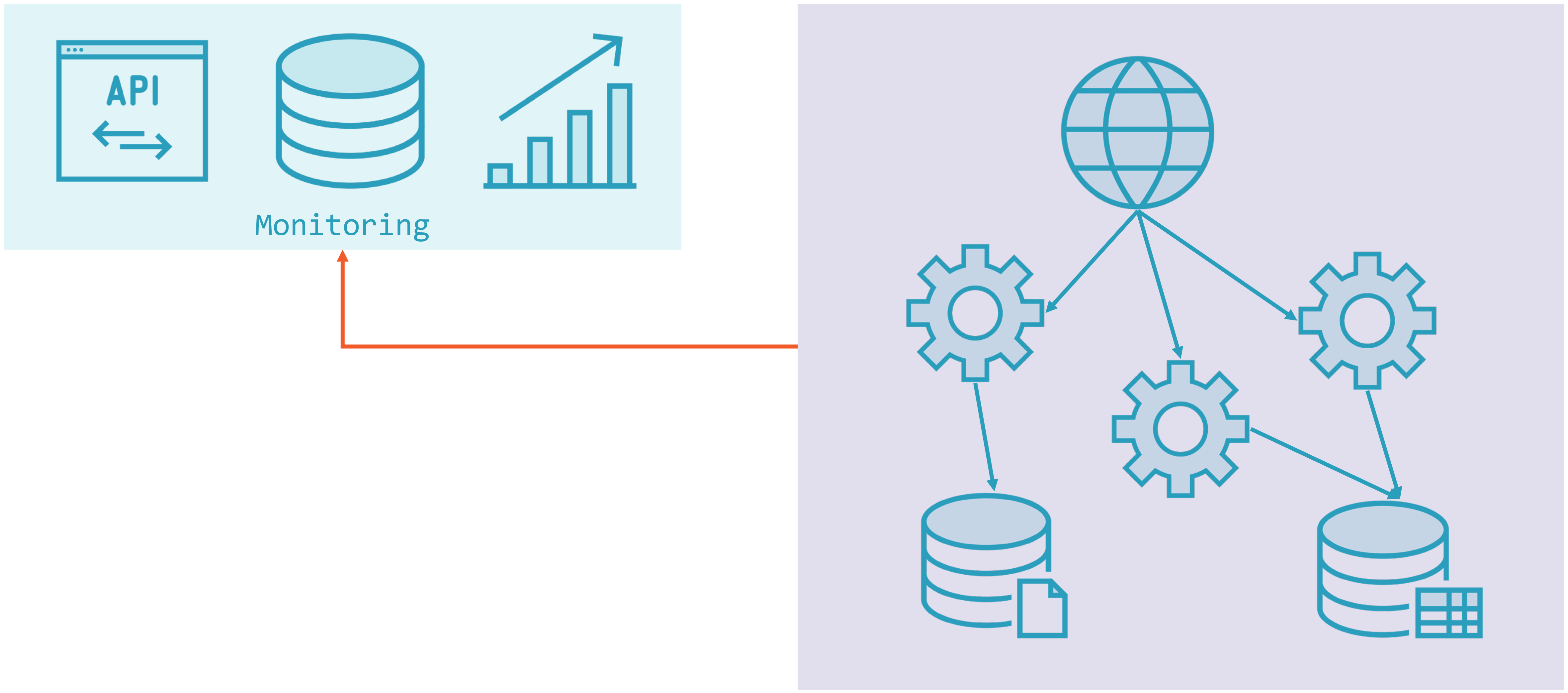




- Health dashboards
- Metric metadata
- Quick triage



Supporting Triage with High-level Metadata


API
Monitoring

Triage: First Responses

Scale up

If compute-bound

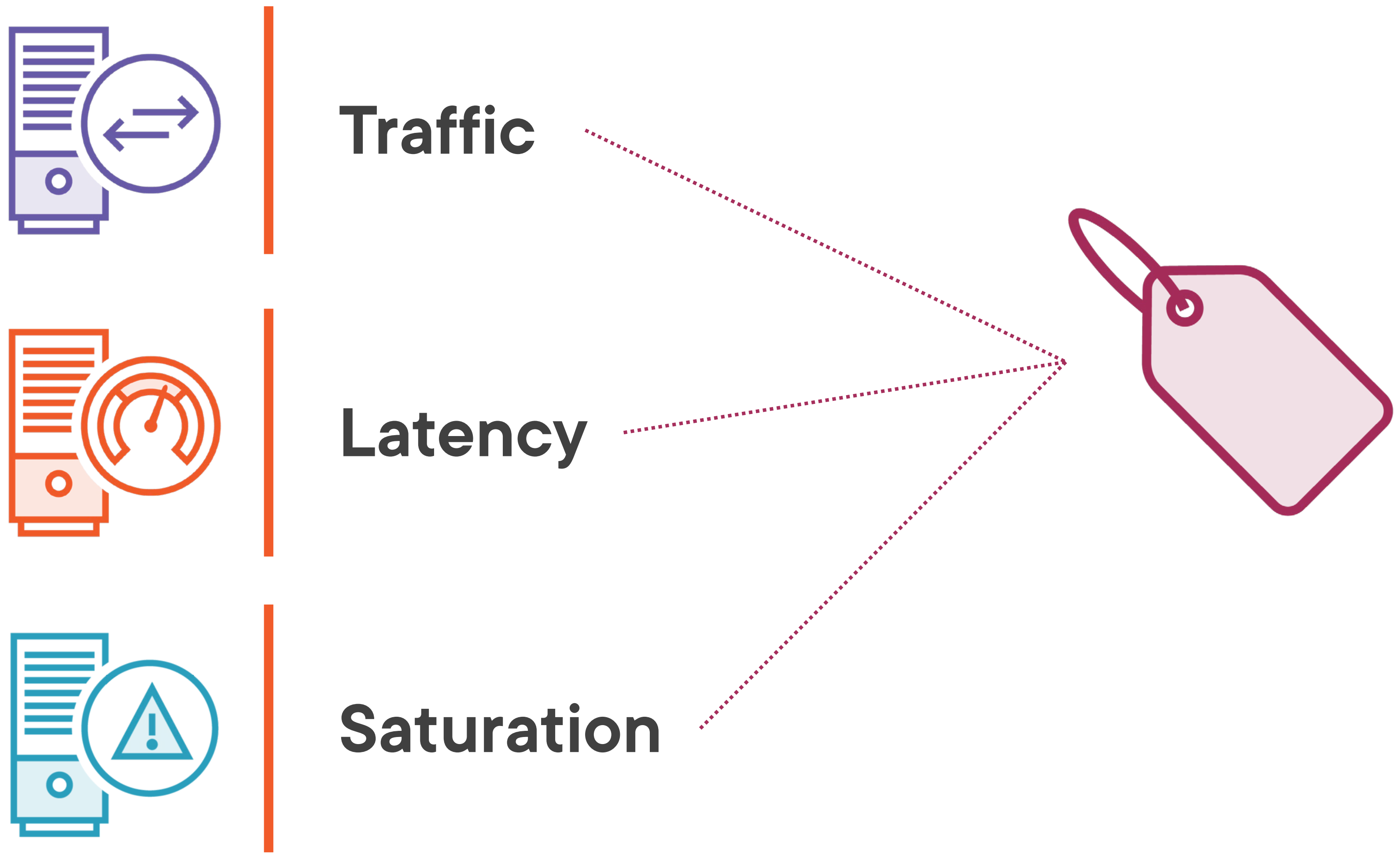
Re-route traffic

If regional

Downgrade service

If global

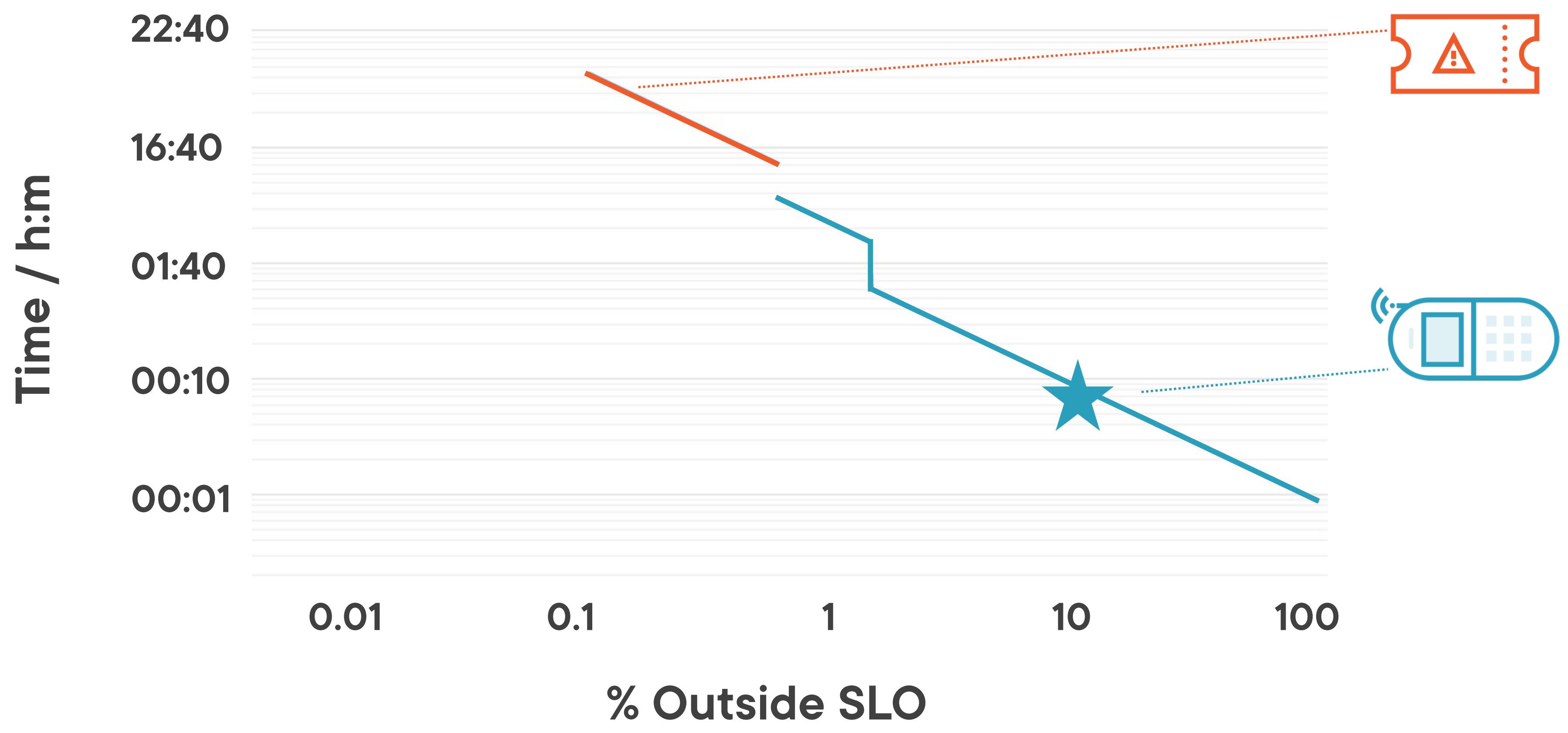
Core Metrics



- Region
- Component

Alerting on Error Budget Burn

Latency **SLO: 99.9% within 2sec**

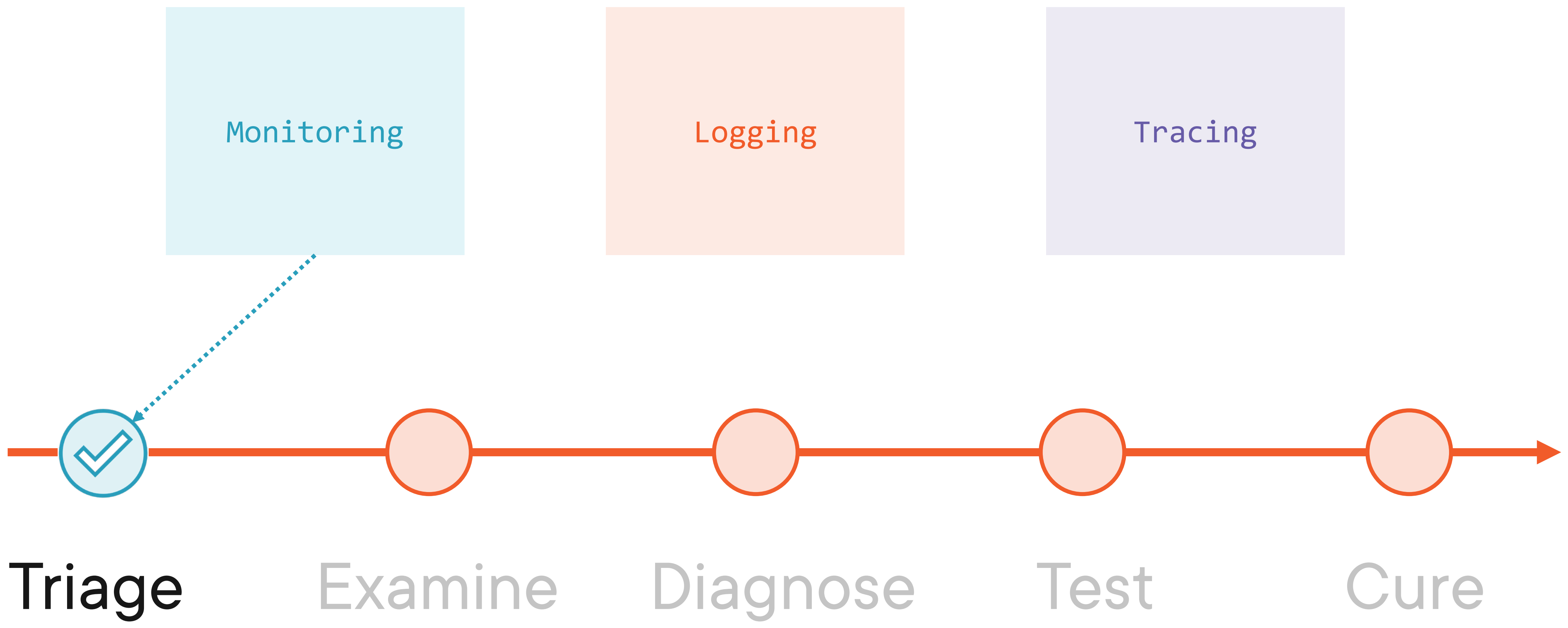


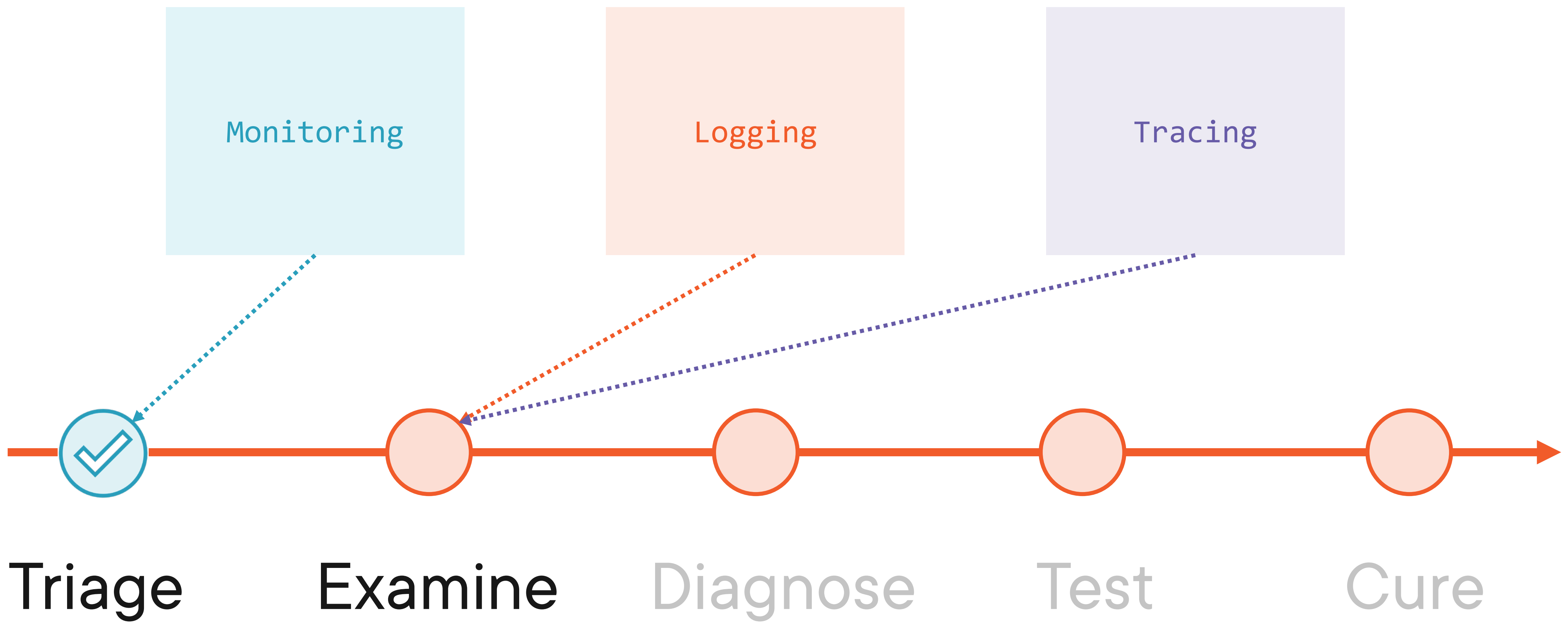
Identifying First Response

Traffic and saturation high globally

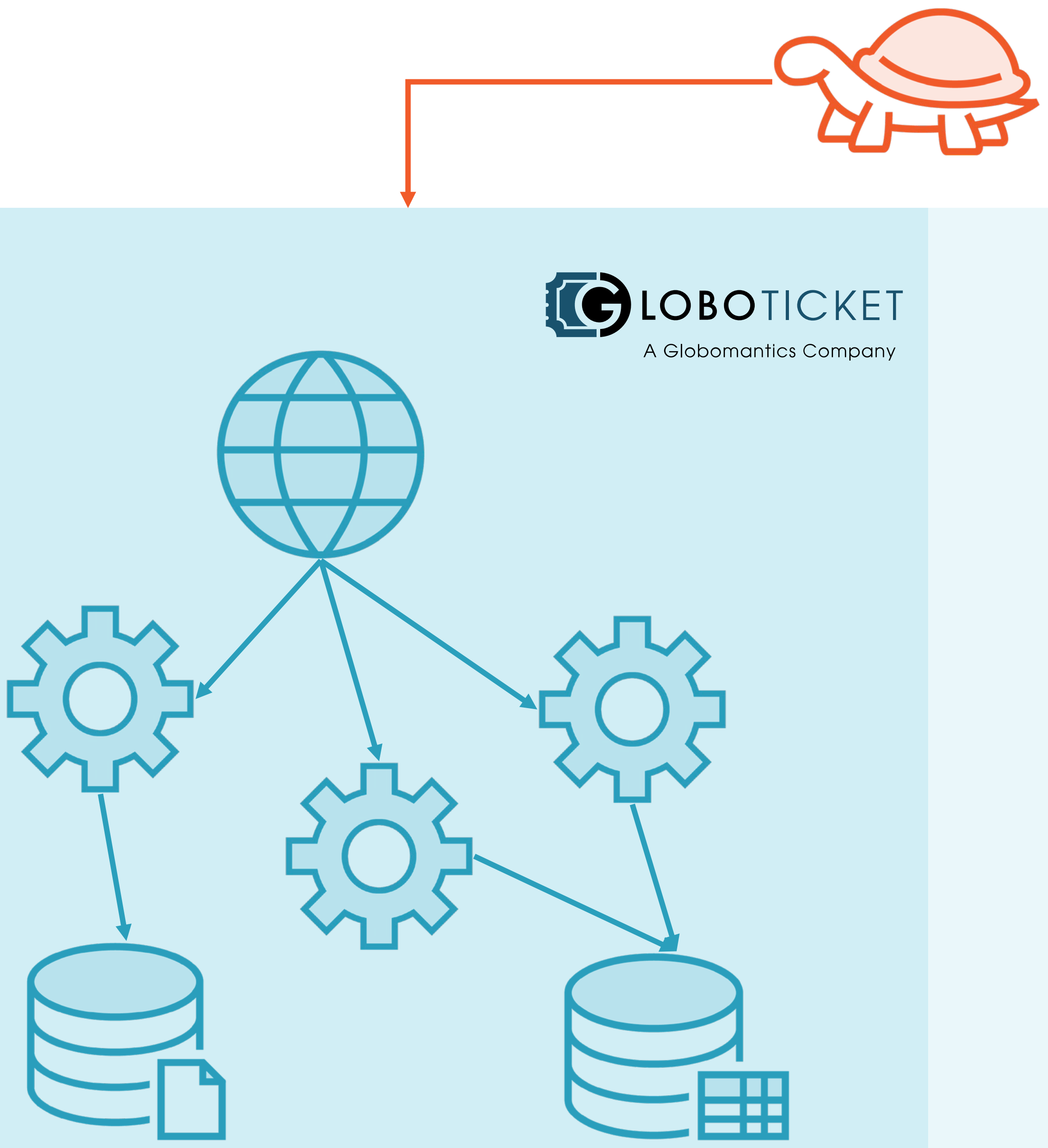
Traffic high for all components in one region

One component saturated in one region

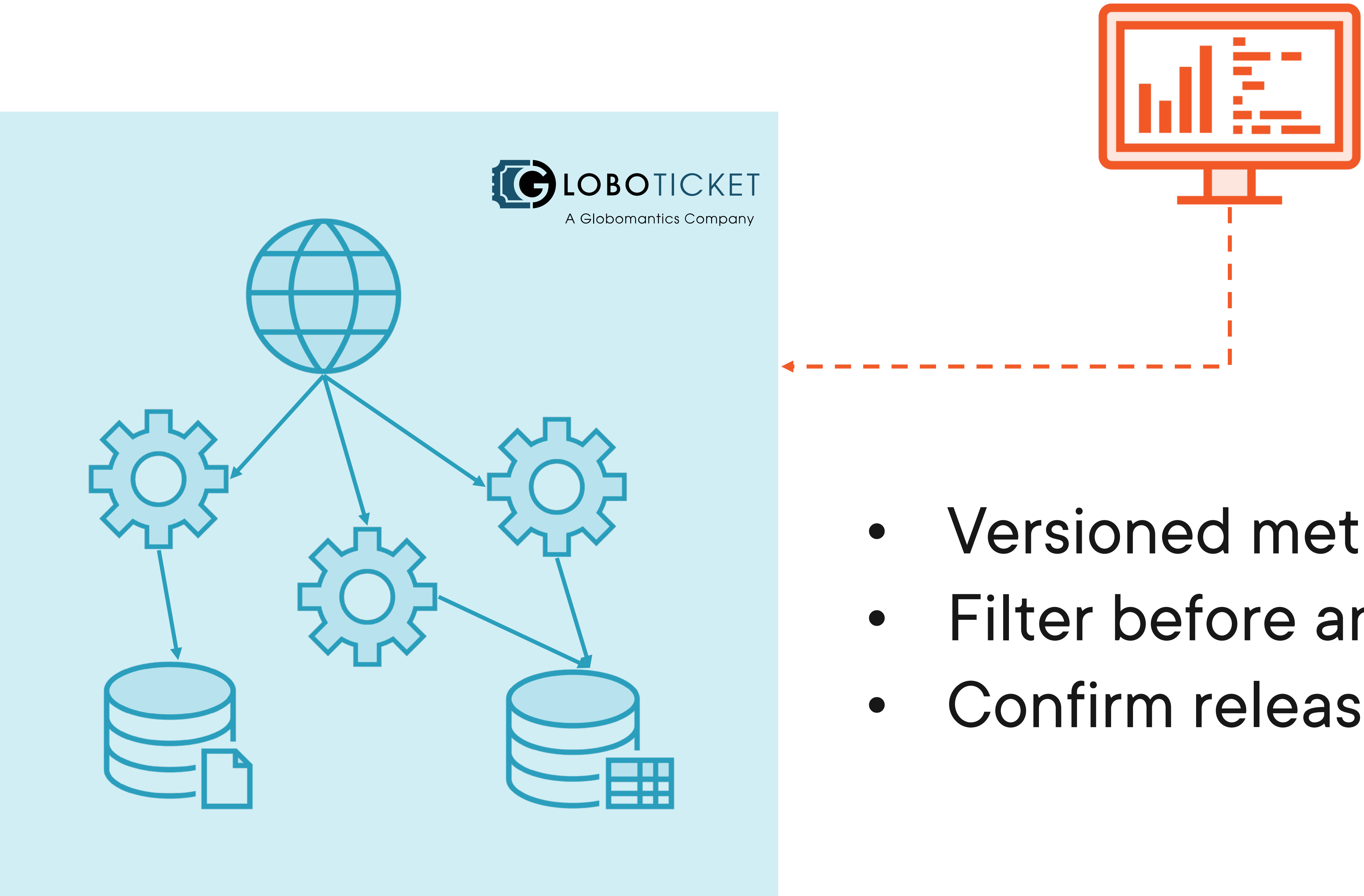

Monitoring
Logging
Tracing
Triage
Examine
Diagnose
Test
Cure


Monitoring
Logging
Tracing
Triage
Examine
Diagnose
Test
Cure

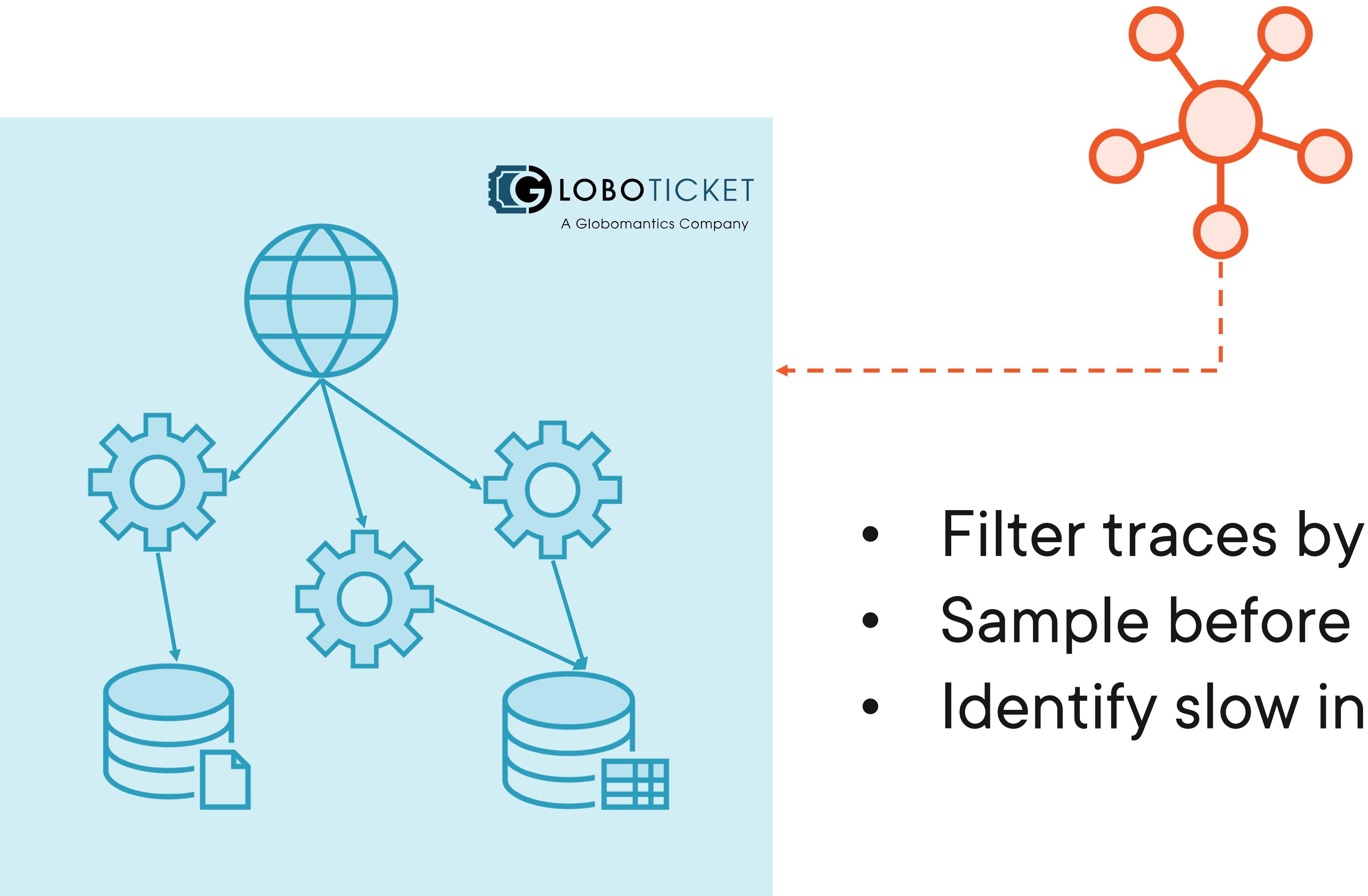
Scenario: Putting Traces to Use



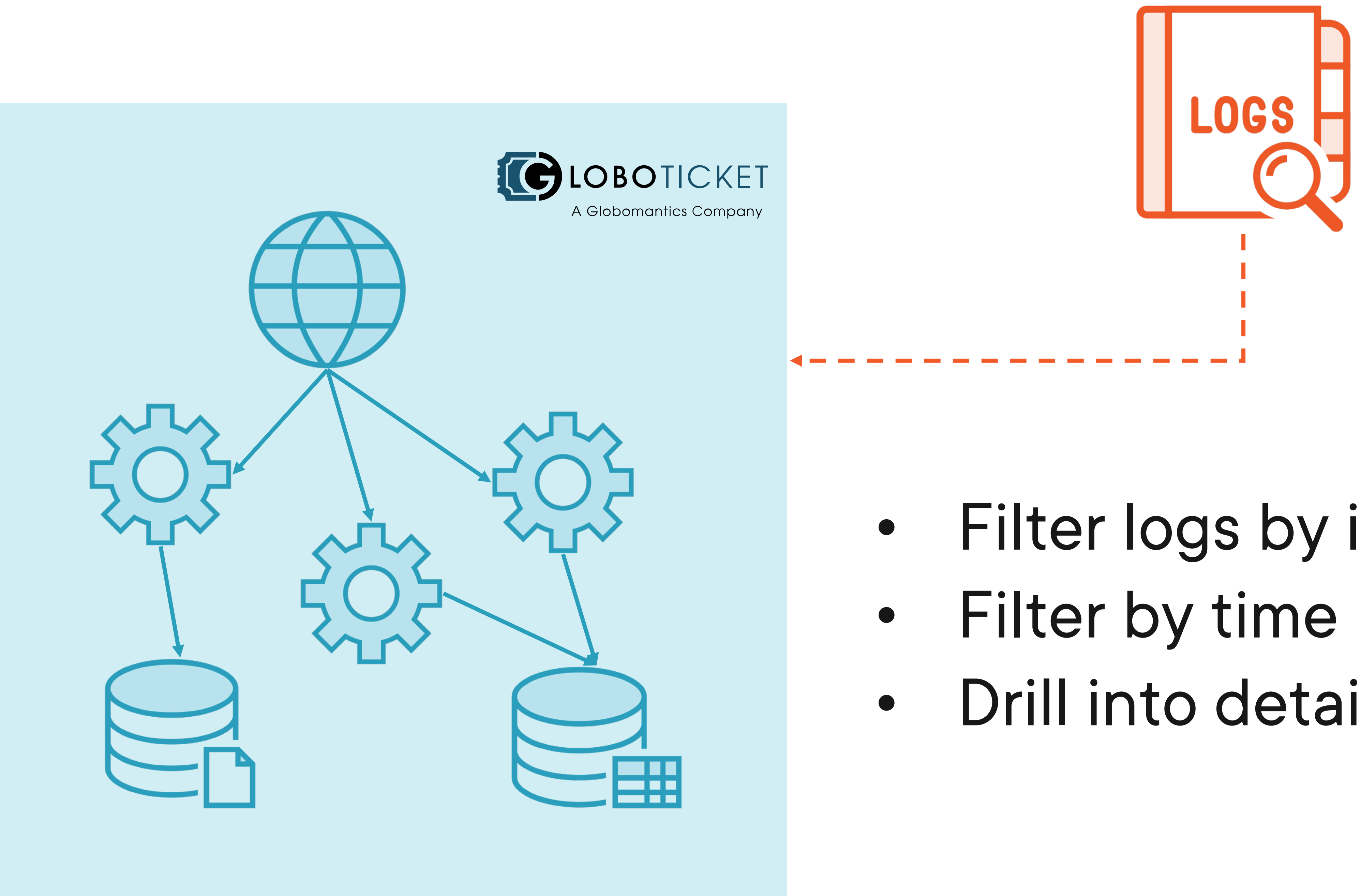
Latency spikes - which is the slow service?



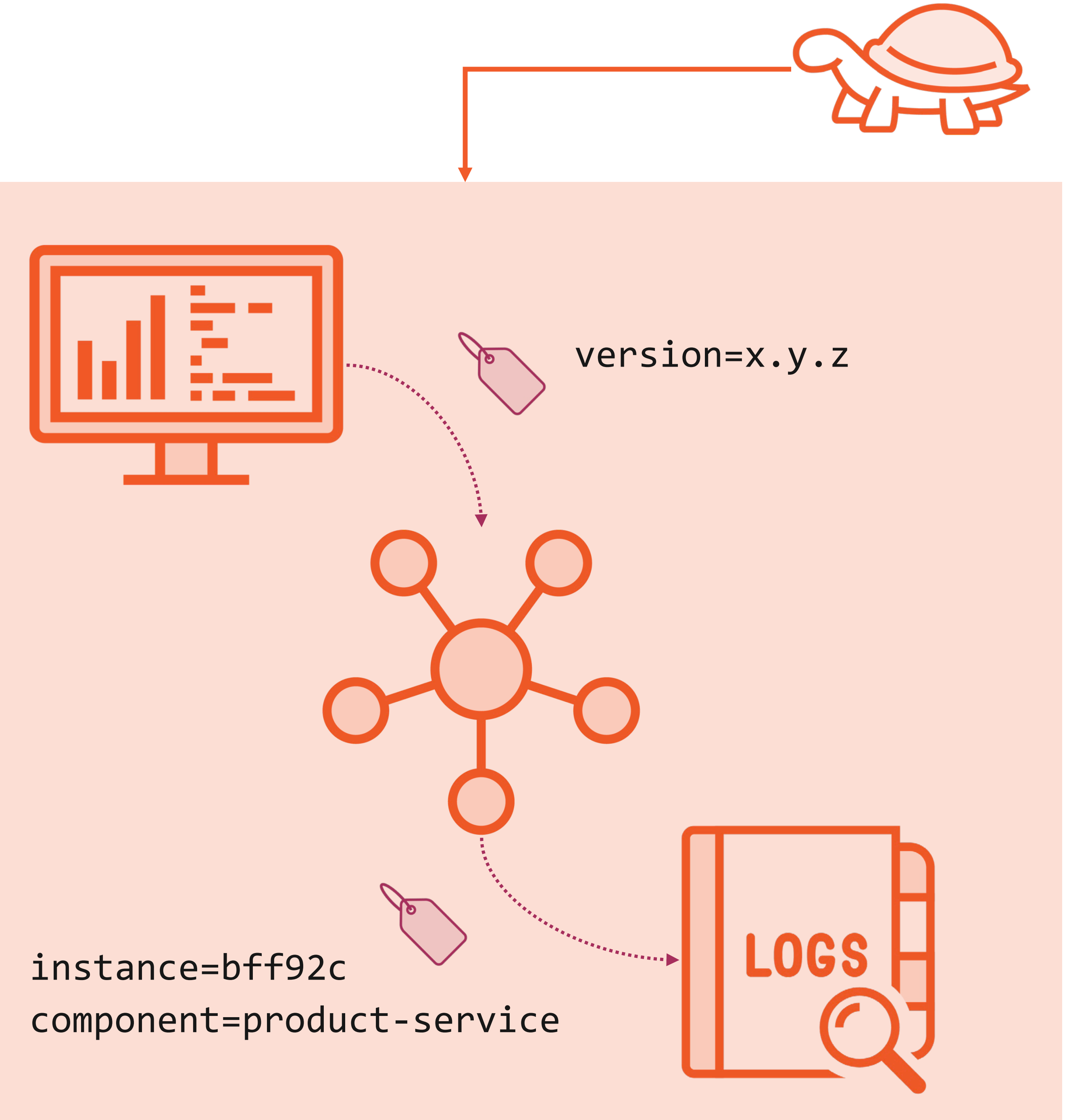
- Versioned metrics
- Filter before and after
- Confirm release trigger



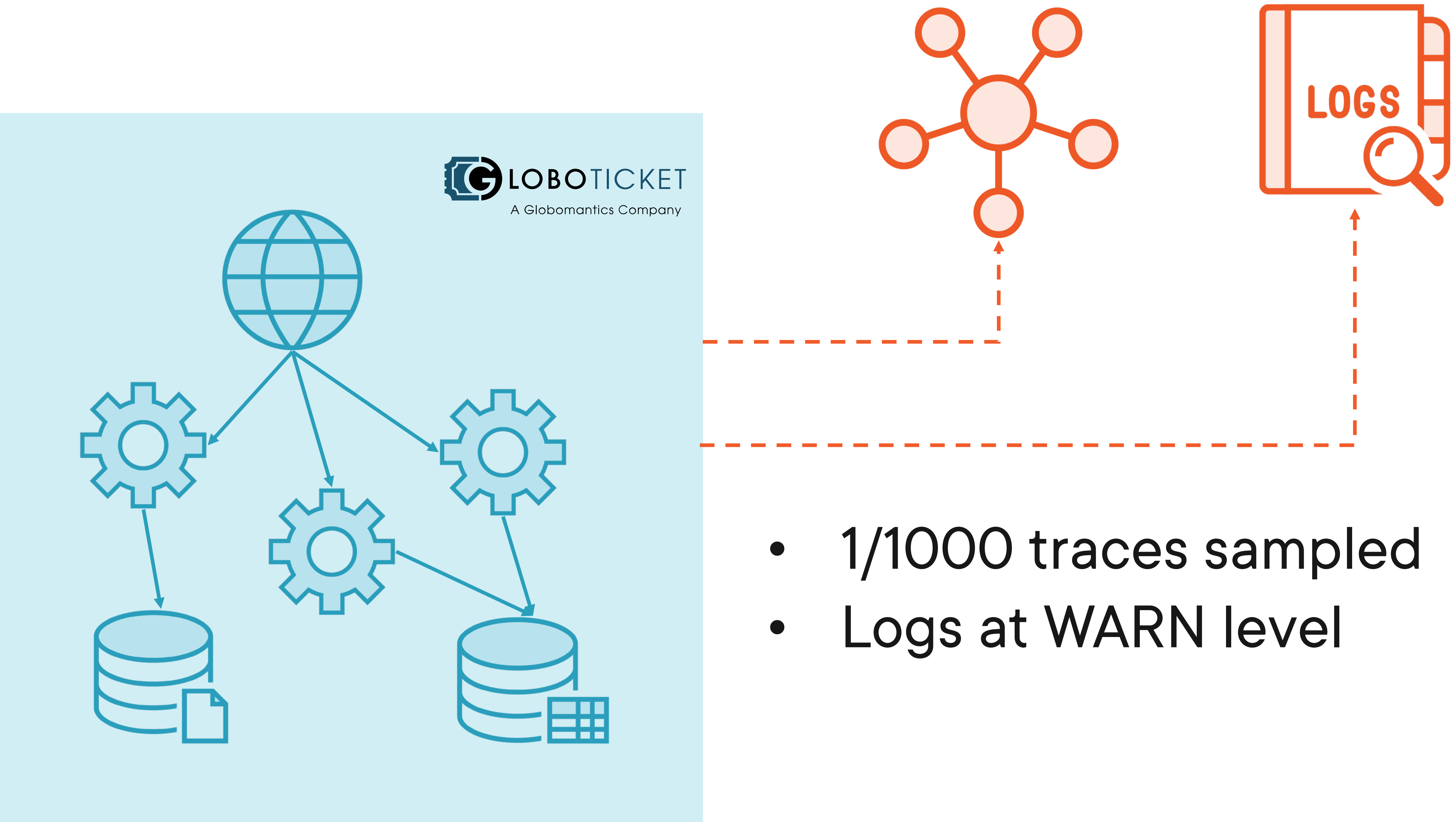
- Filter traces by time
- Sample before and after
- Identify slow instance



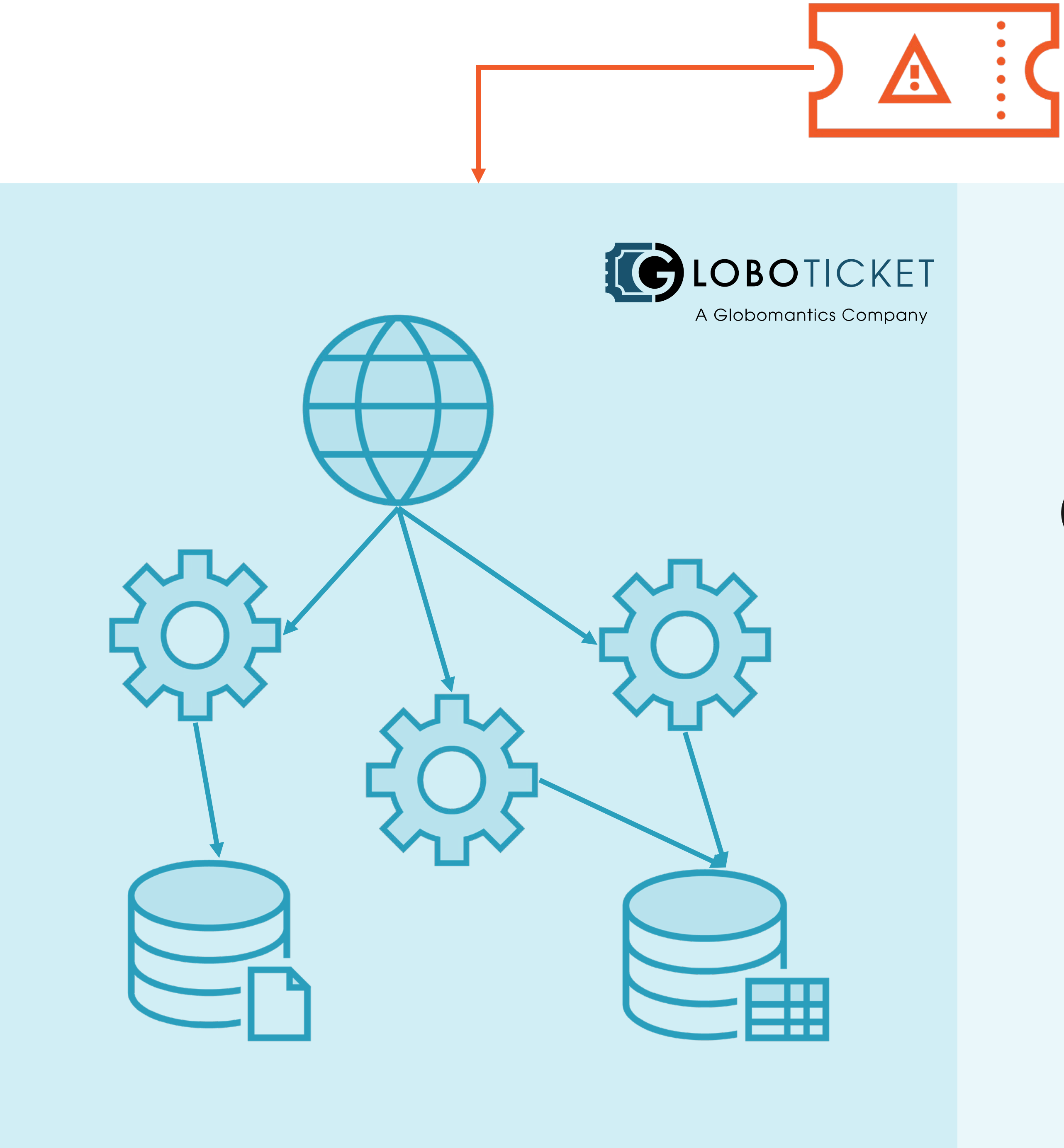
- Filter logs by instance
- Filter by time
- Drill into detail



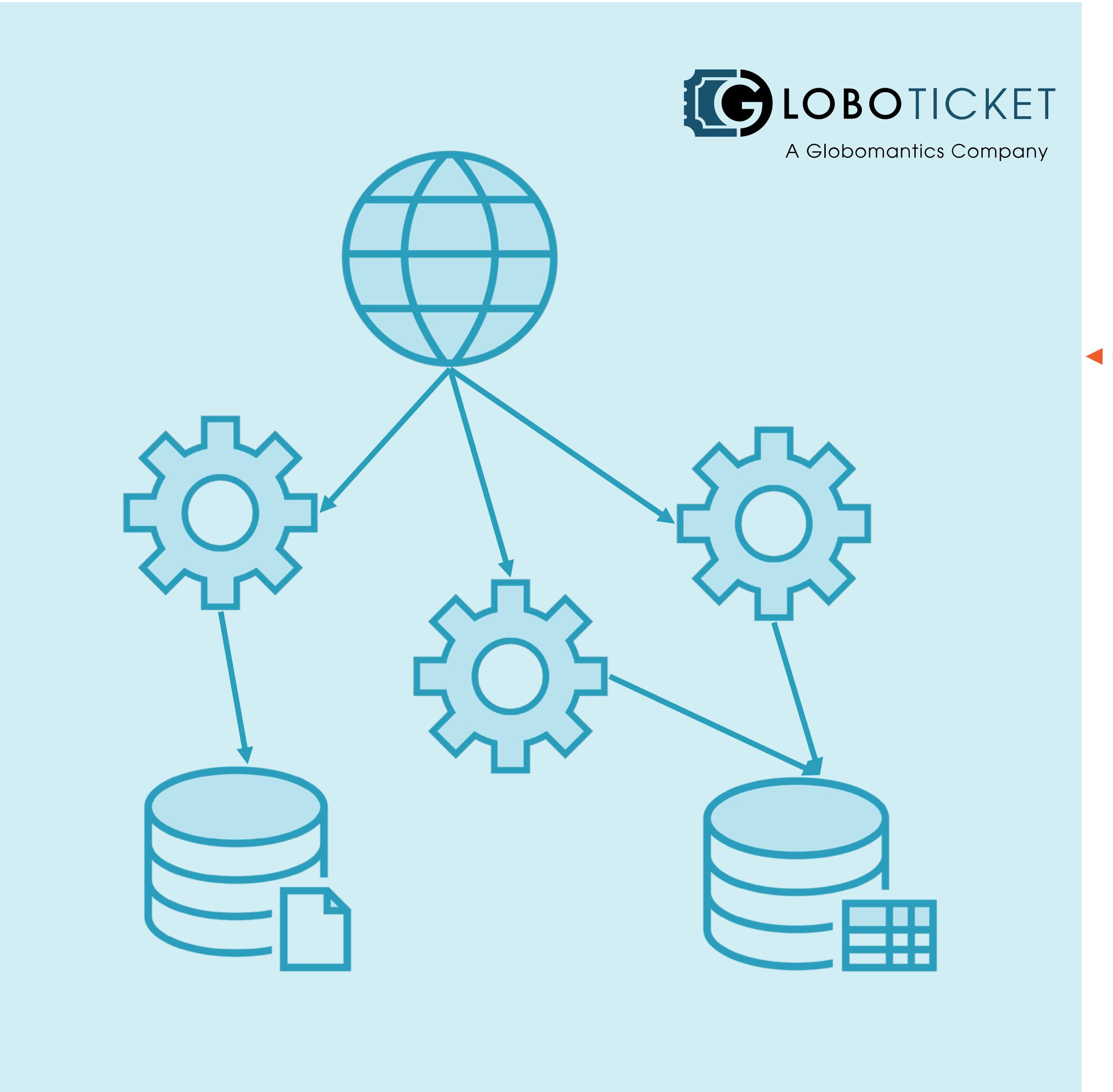
- Version in traces
- Trace ID in logs
- Is the data there?



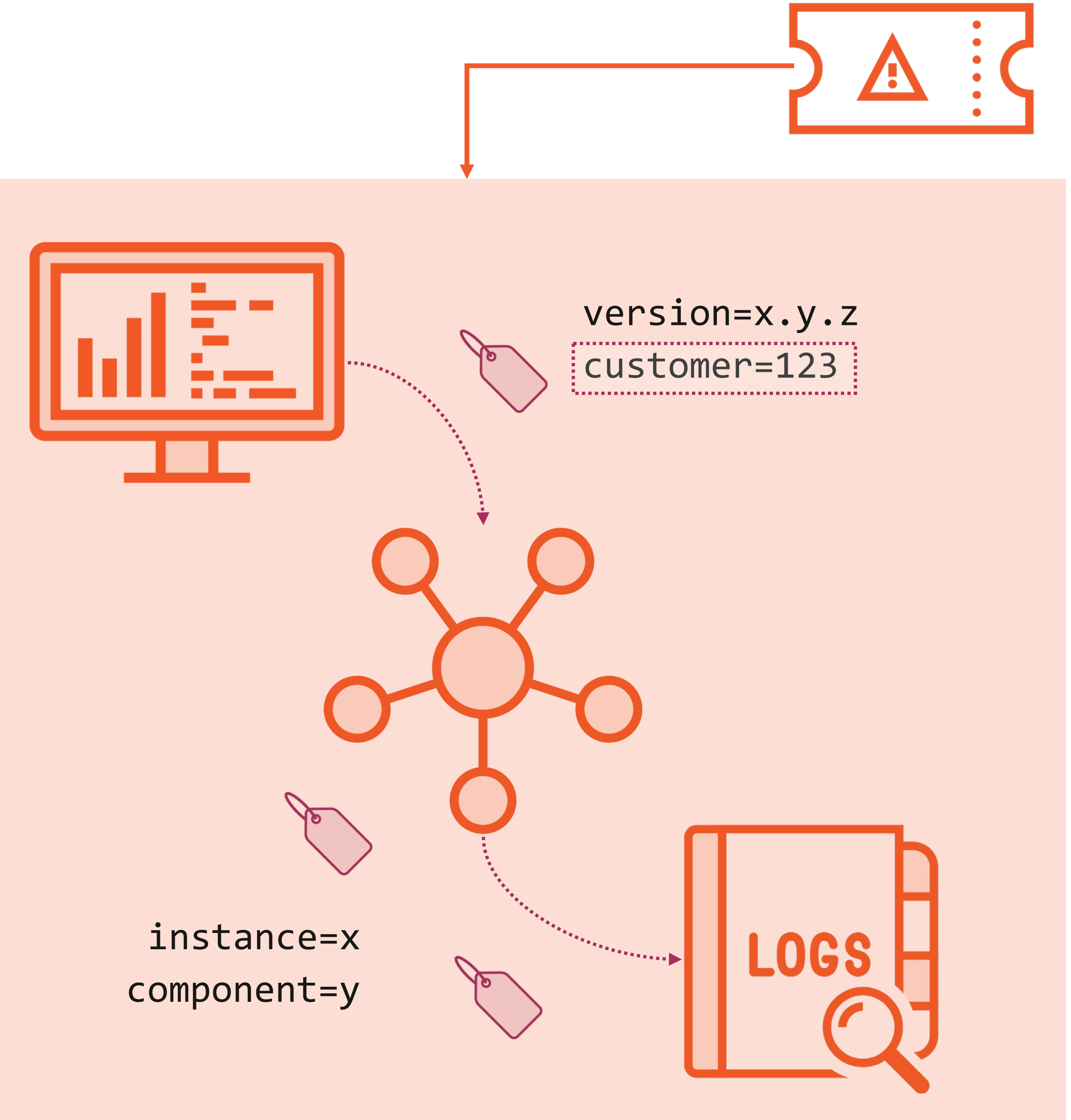
- 1/1000 traces sampled
- Logs at WARN level



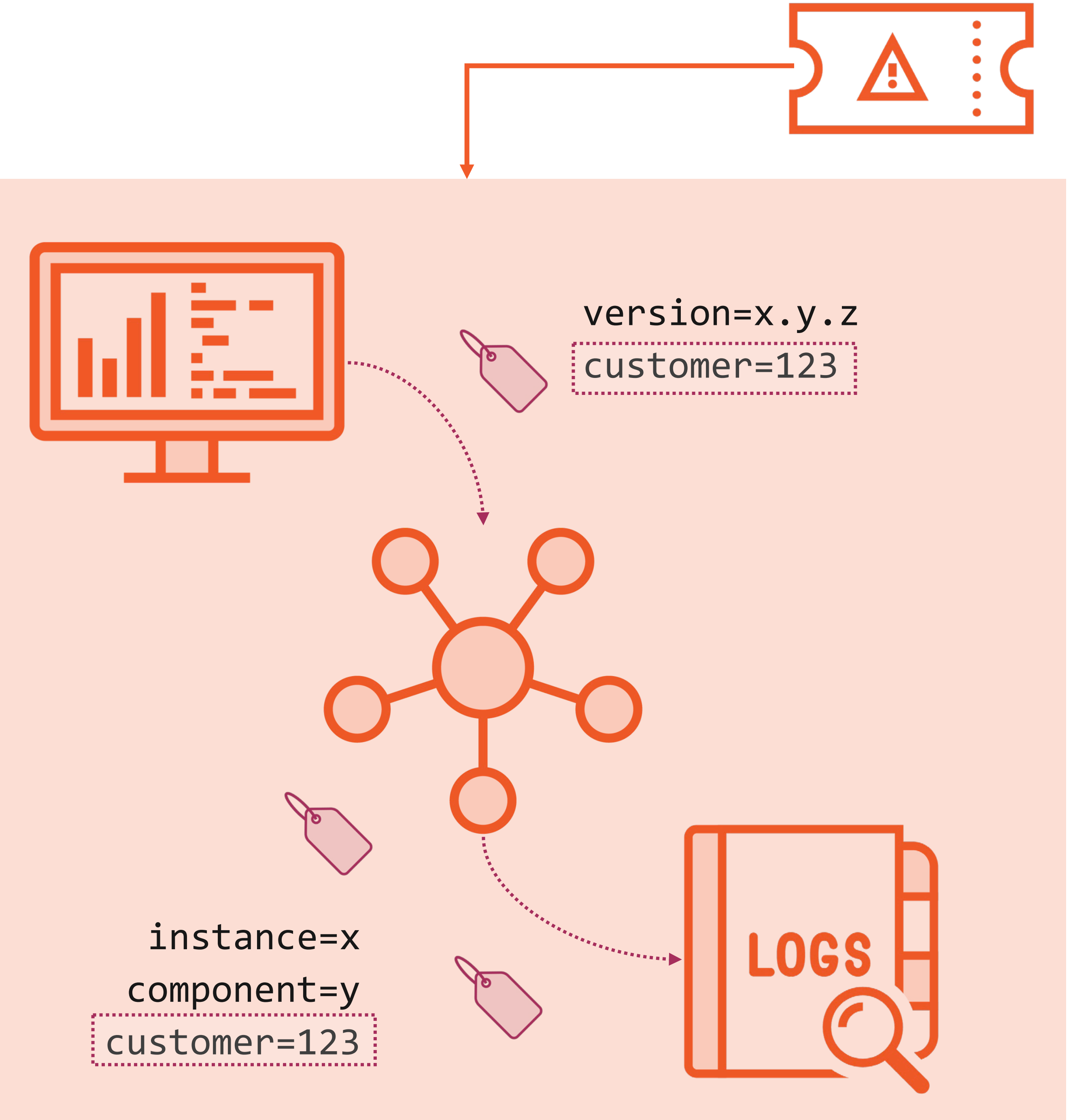
Poor experience for one customer - why?



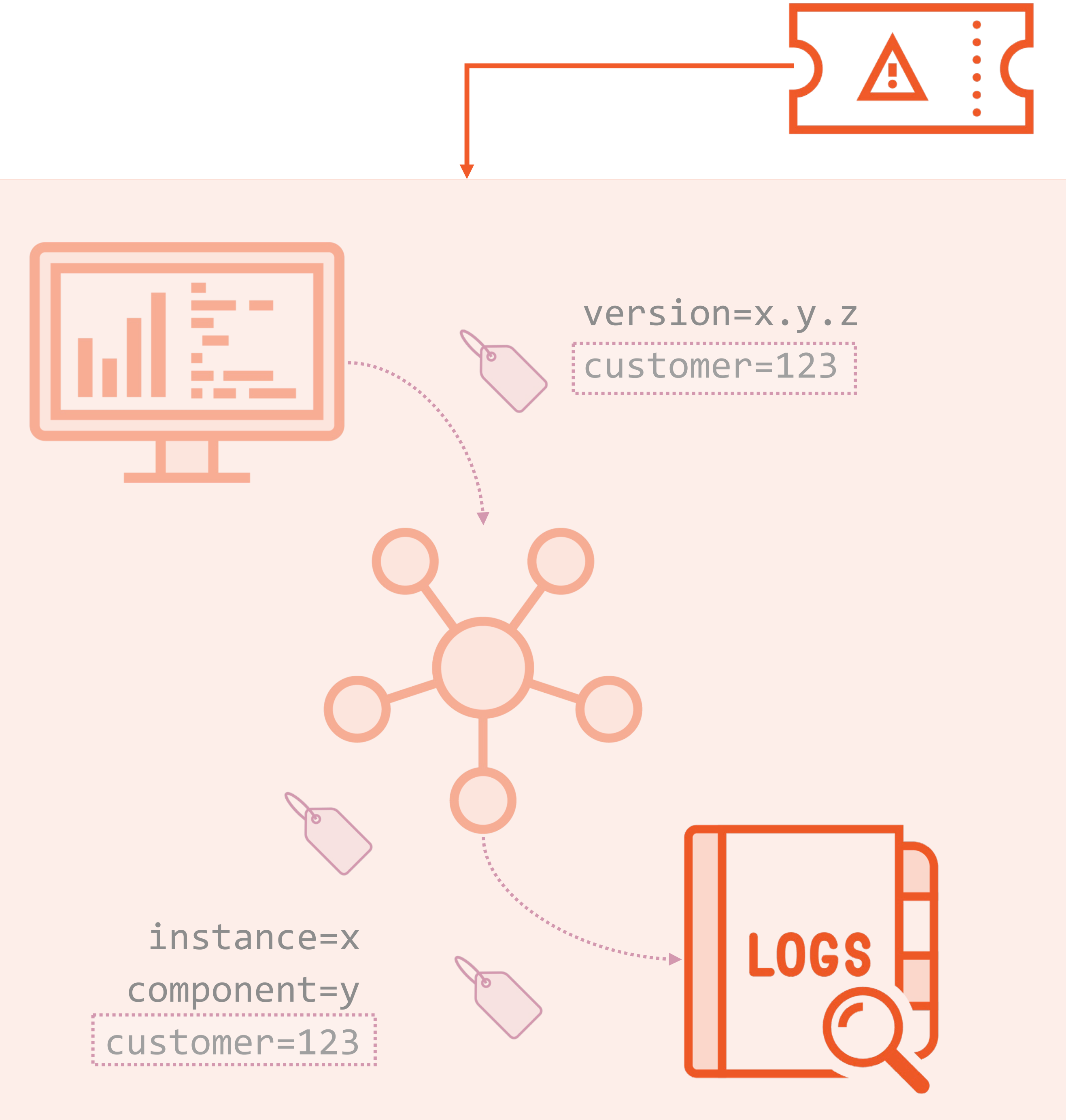
- No customer IDs
- Specific feature?
- Call product team



- Low-level metadata
- Not for all metrics
- Customer SLOs

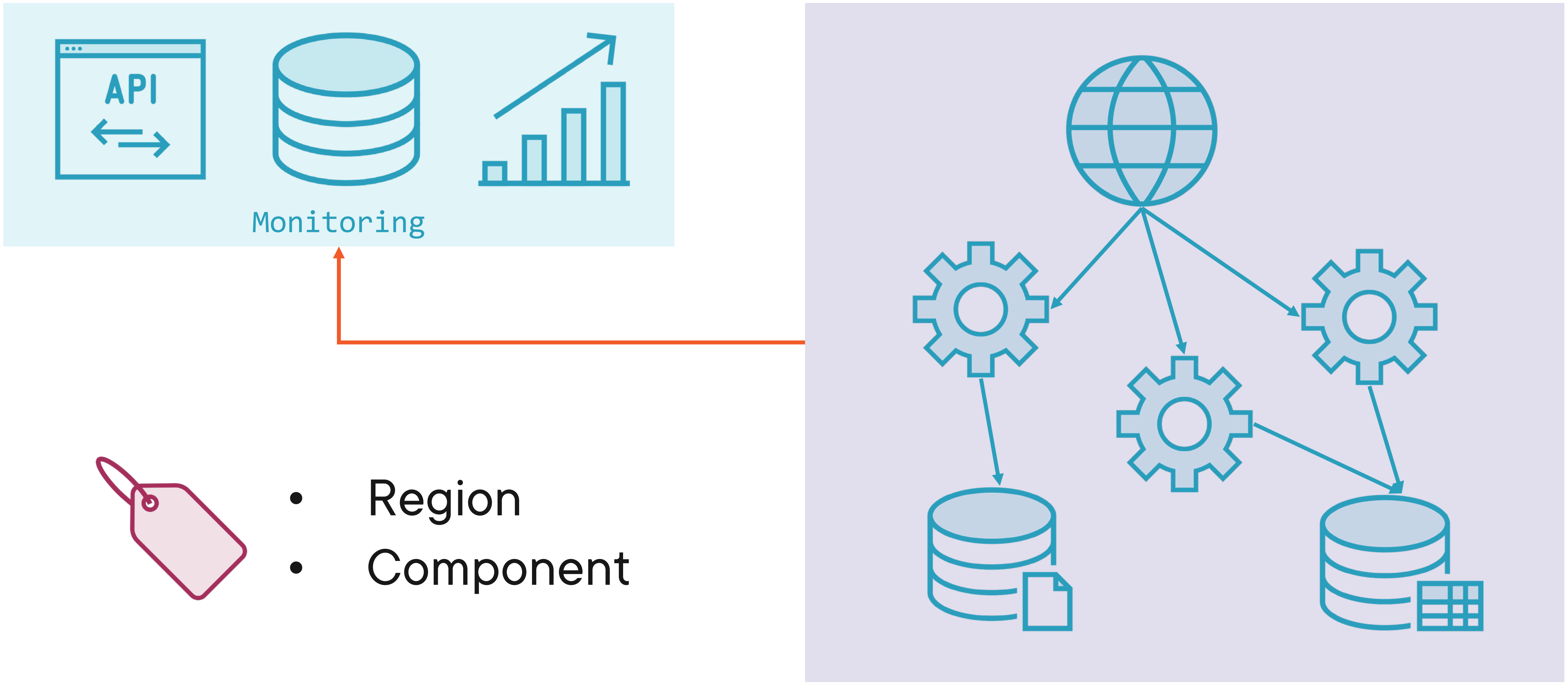


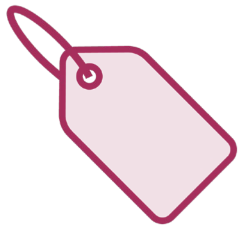
- Trace by customer
- Add ID to logs
- Work backwards



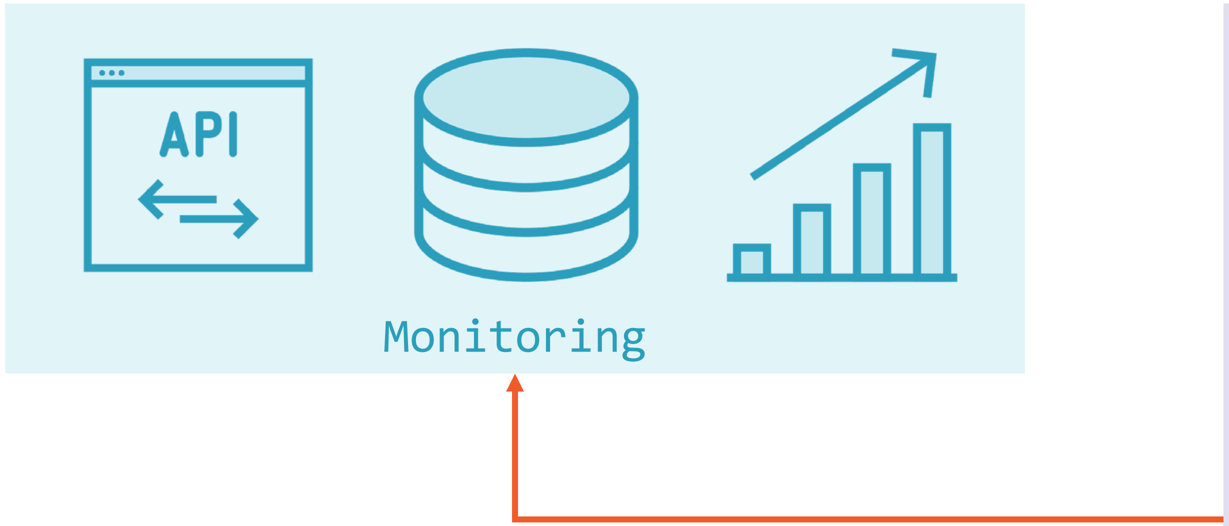
- Is the data there?

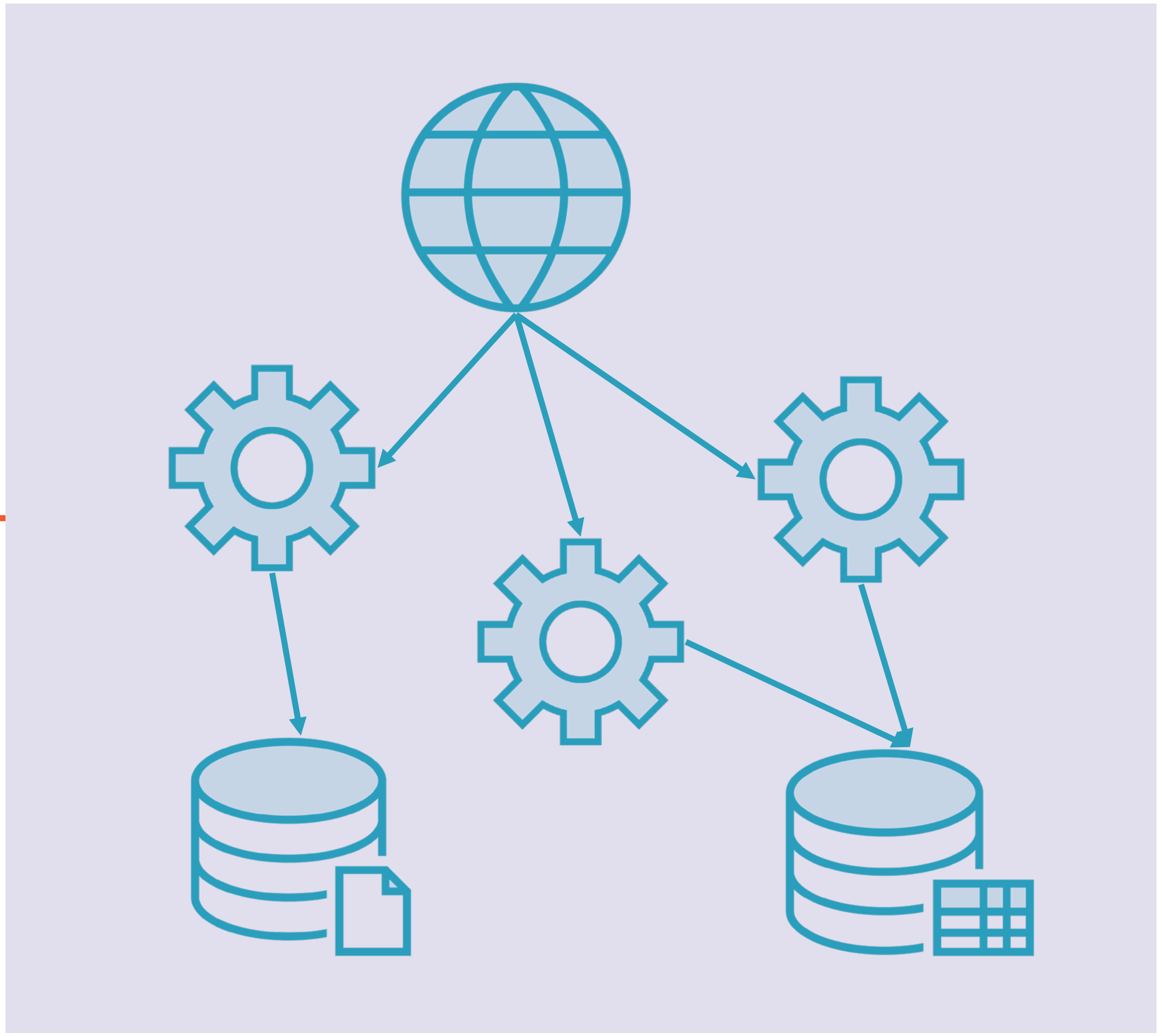
Supporting Examination with Low-level Metadata





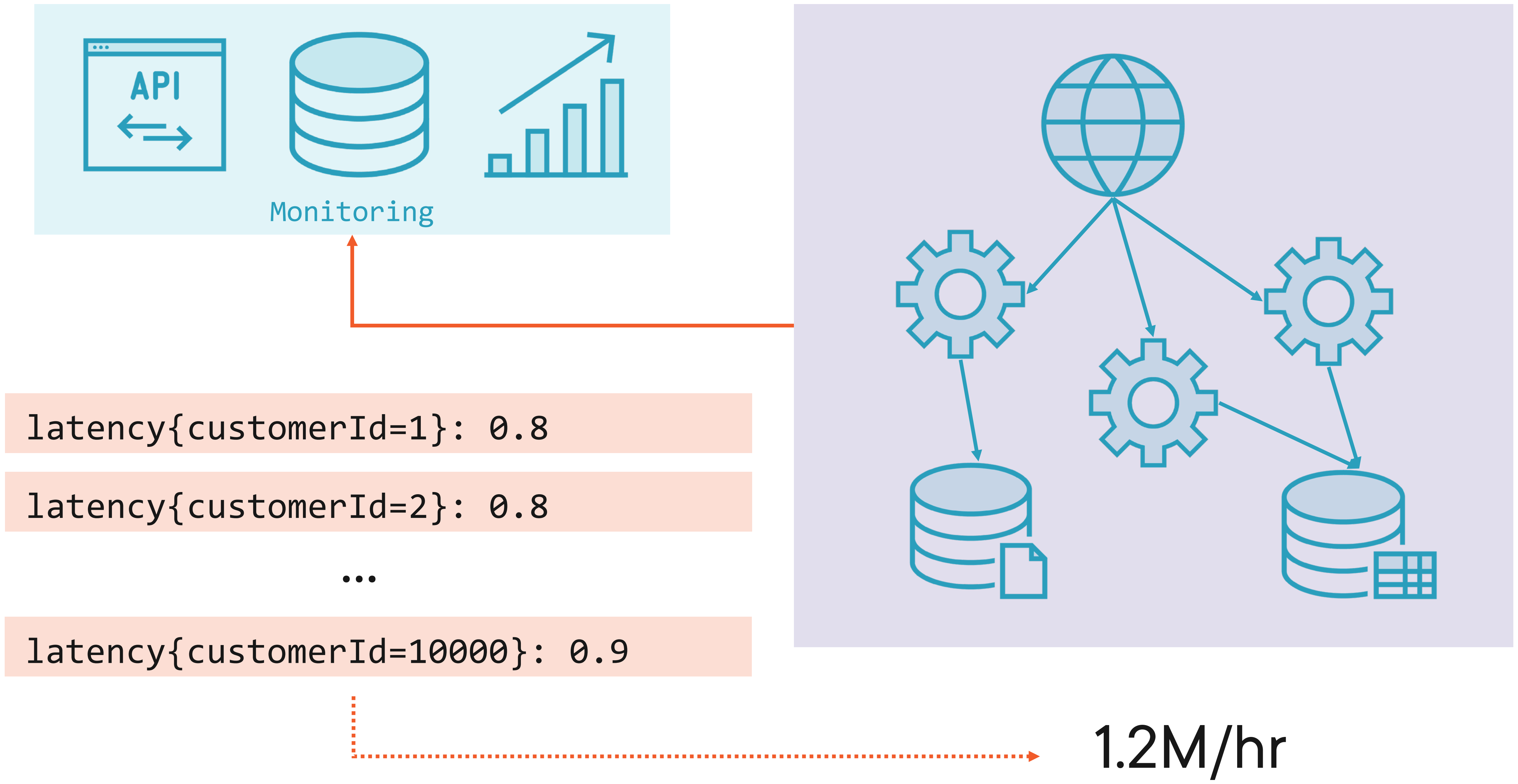
- Region
- Component



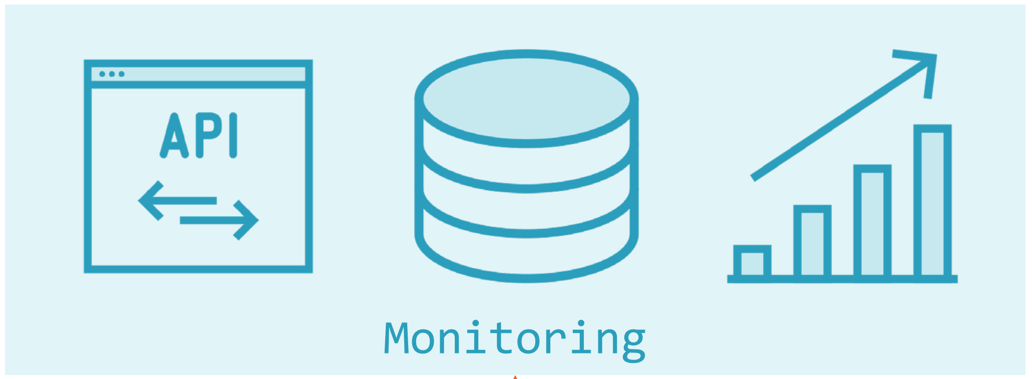


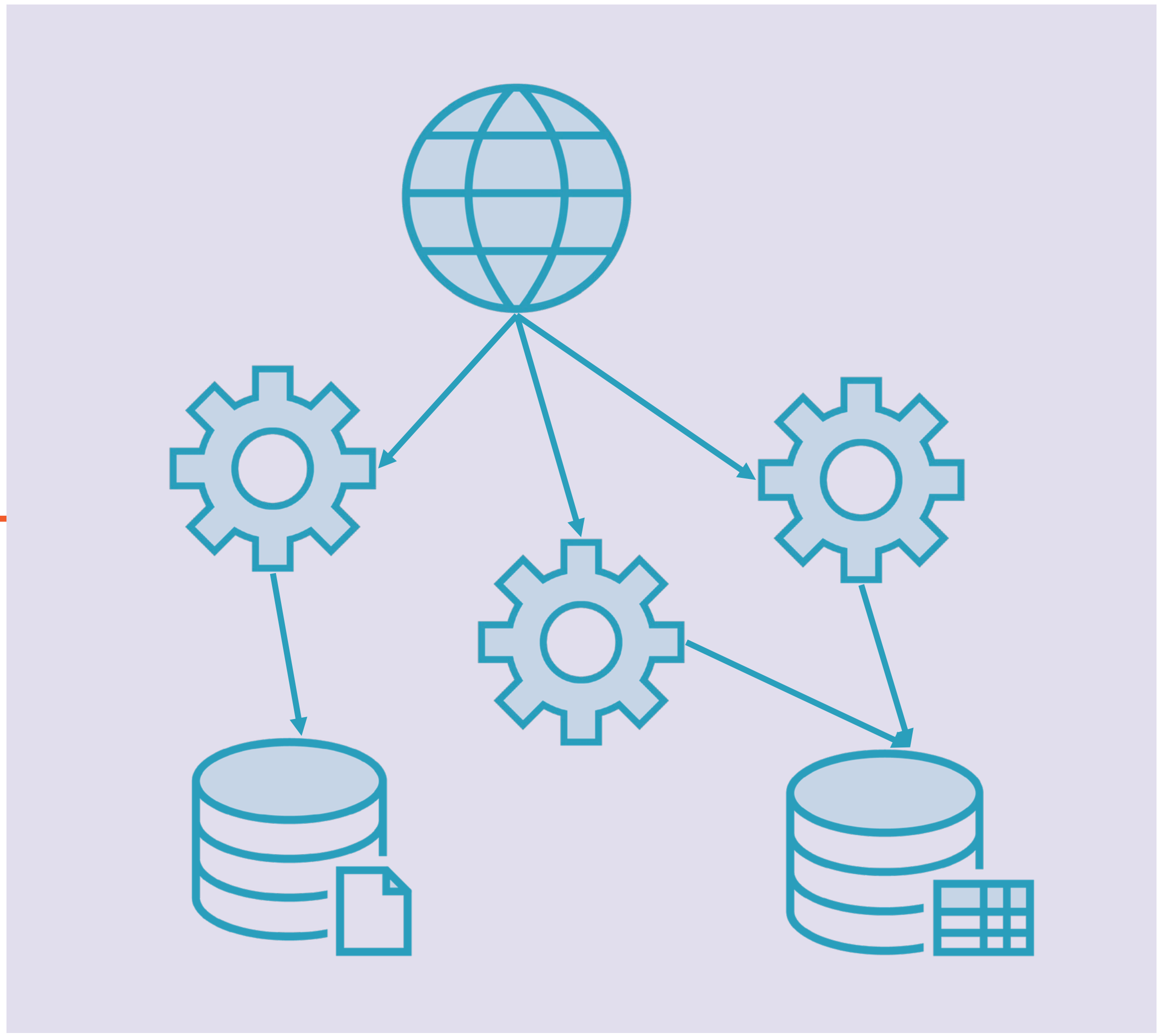
```
latency{component=1,region=1}: 0.9
latency{component=1,region=2}: 0.7
...
latency{component=20,region=4}: 0.1
```

480/hr

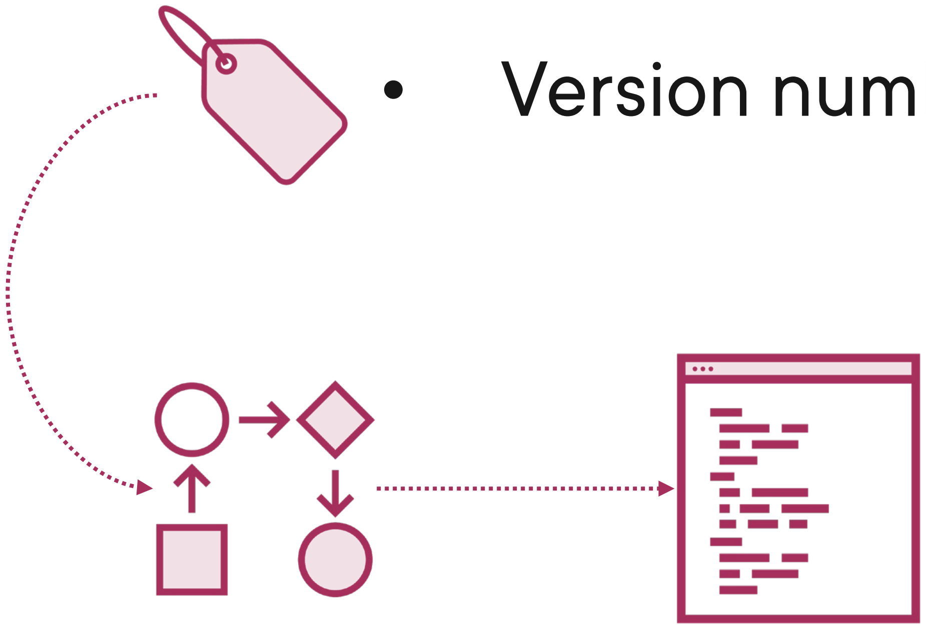


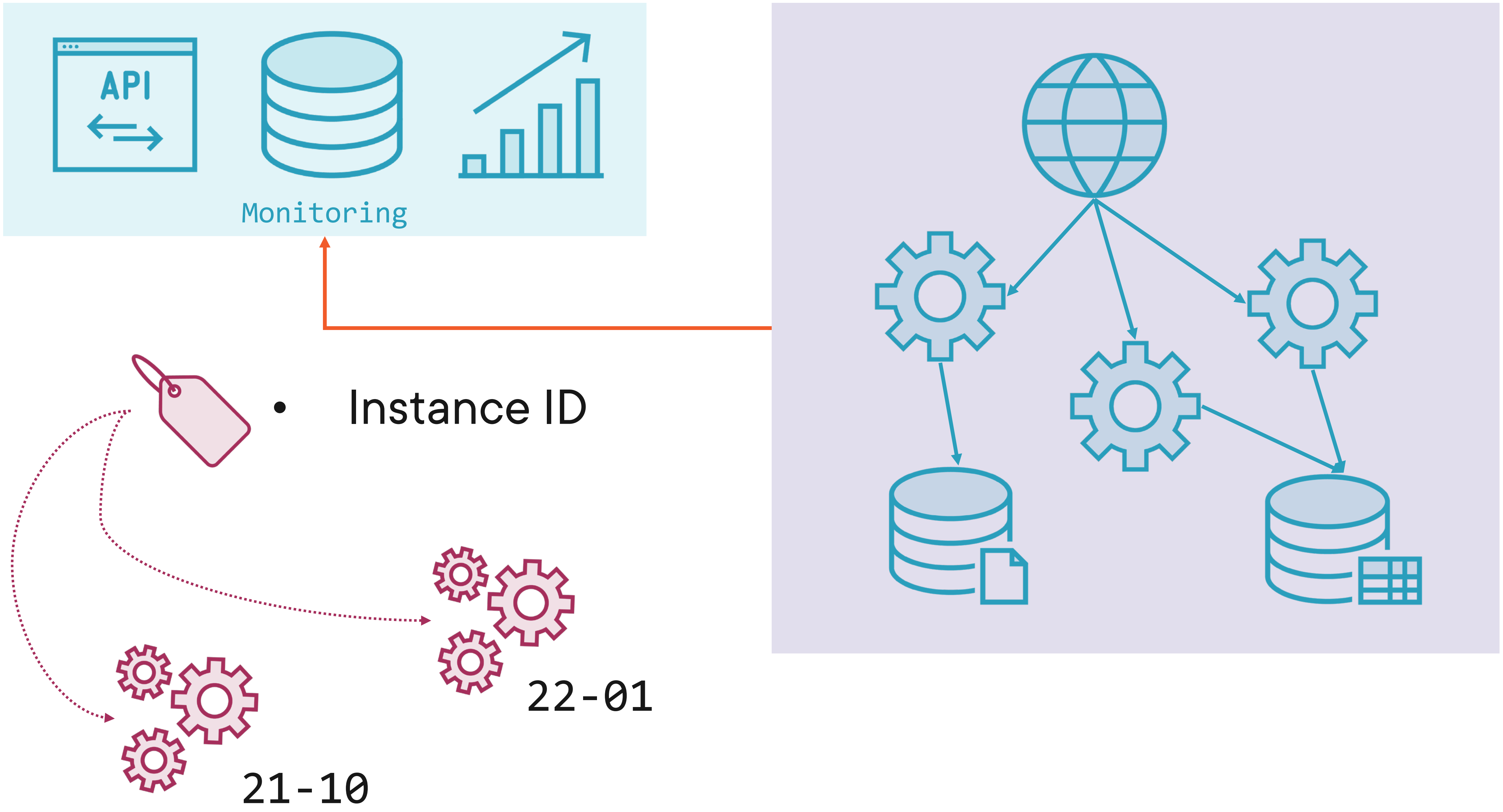
```
latency{customerId=1}: 0.8
latency{customerId=2}: 0.8
...
latency{customerId=10000}: 0.9
```



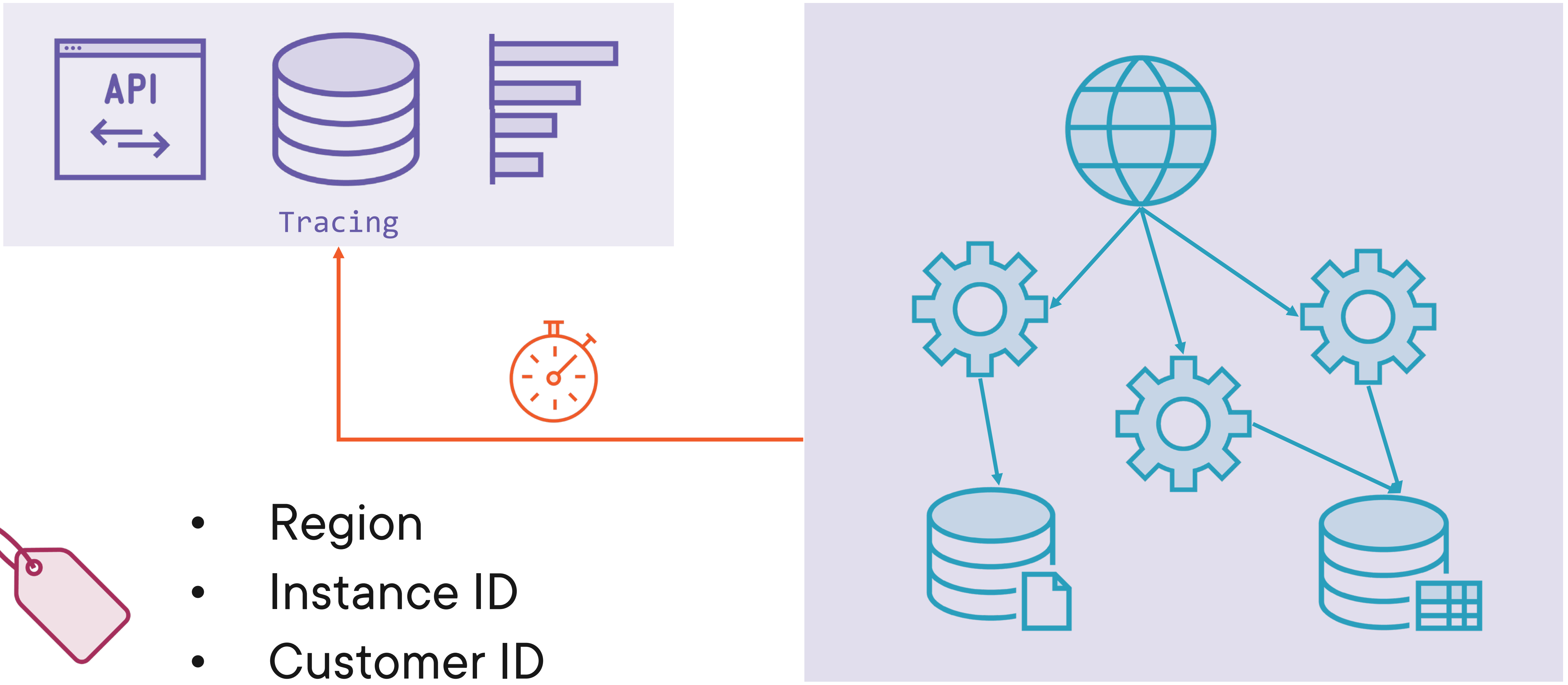


- Version number



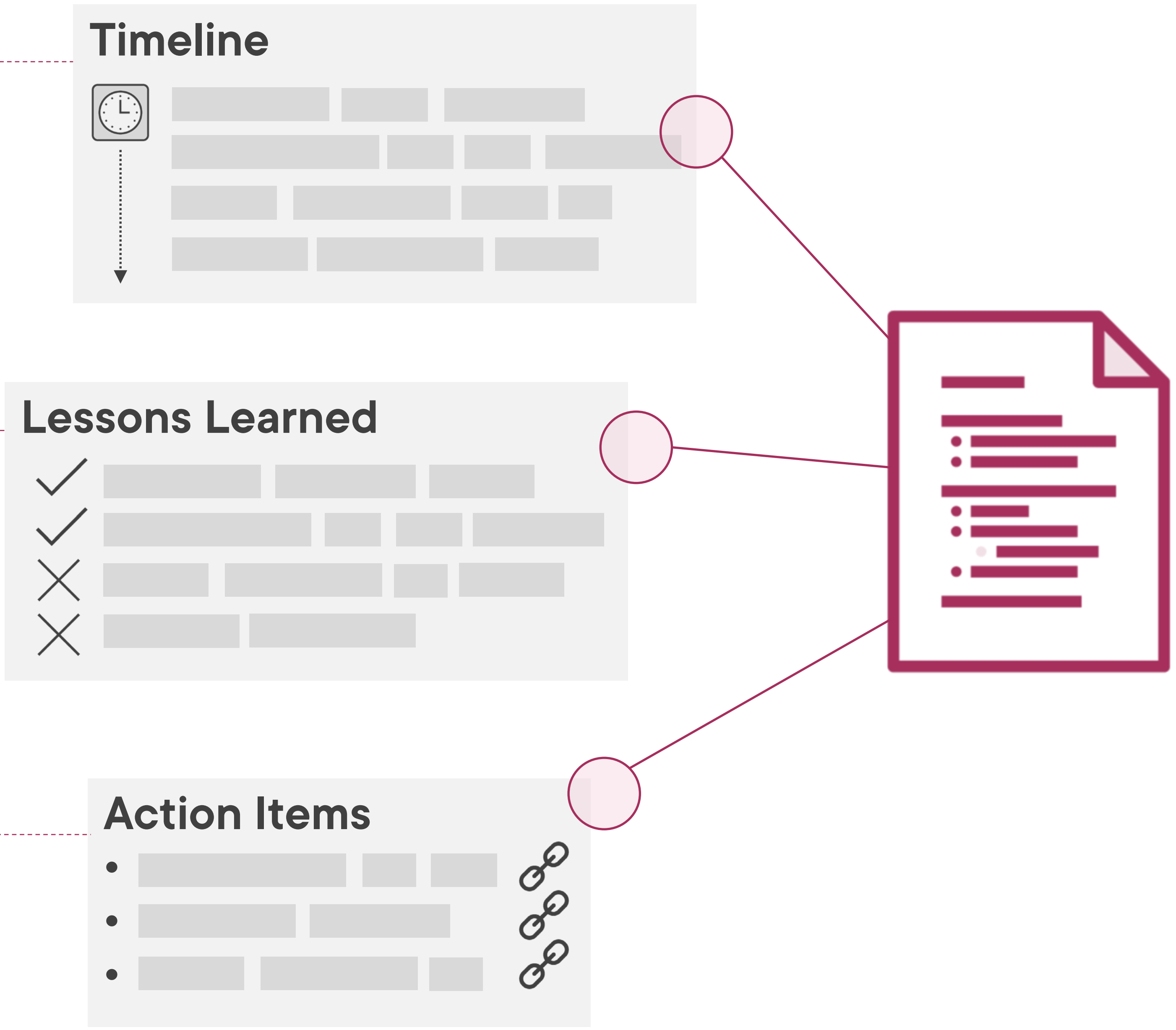


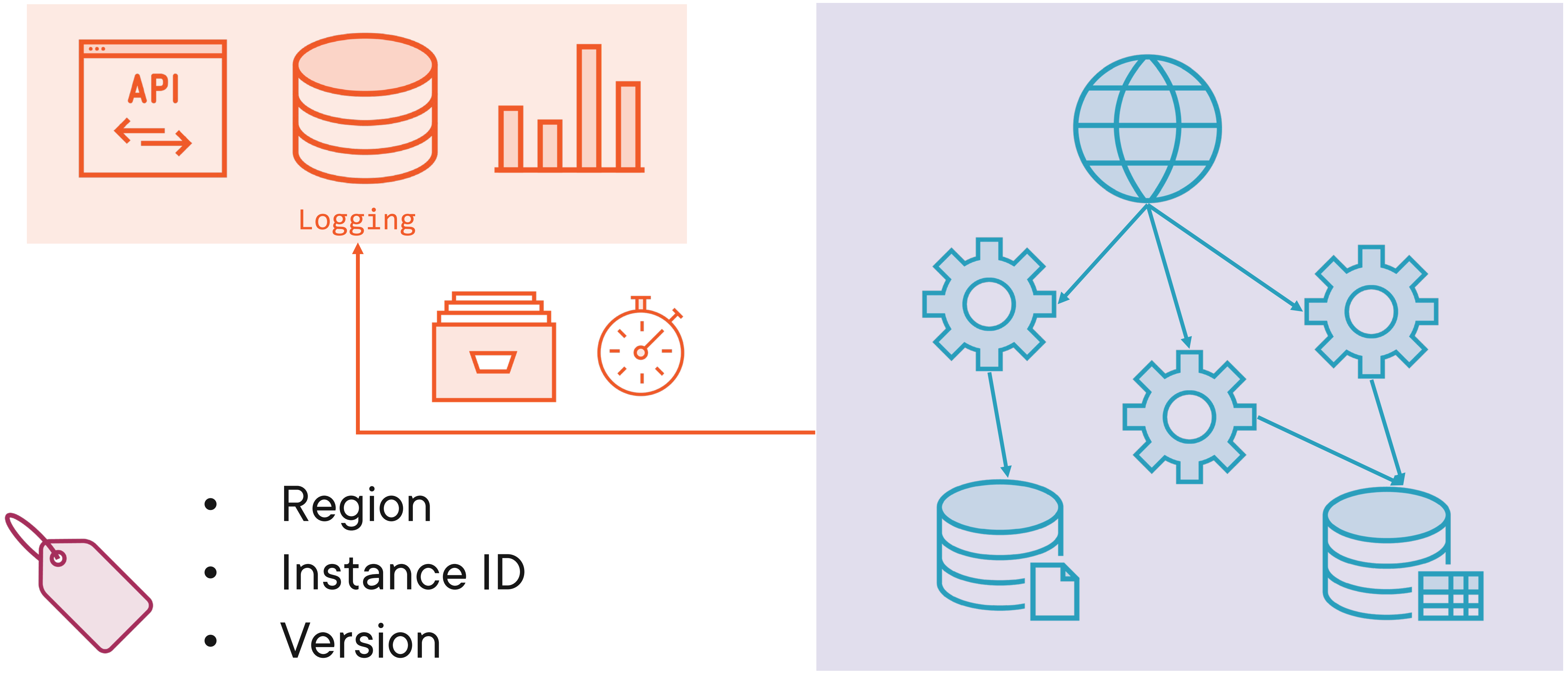
- Instance ID



- Region
- Instance ID
- Customer ID

- Identifying region took >1hr
- Need metadata to correlate metrics & traces
- Add region ID to Prometheus & Jaeger

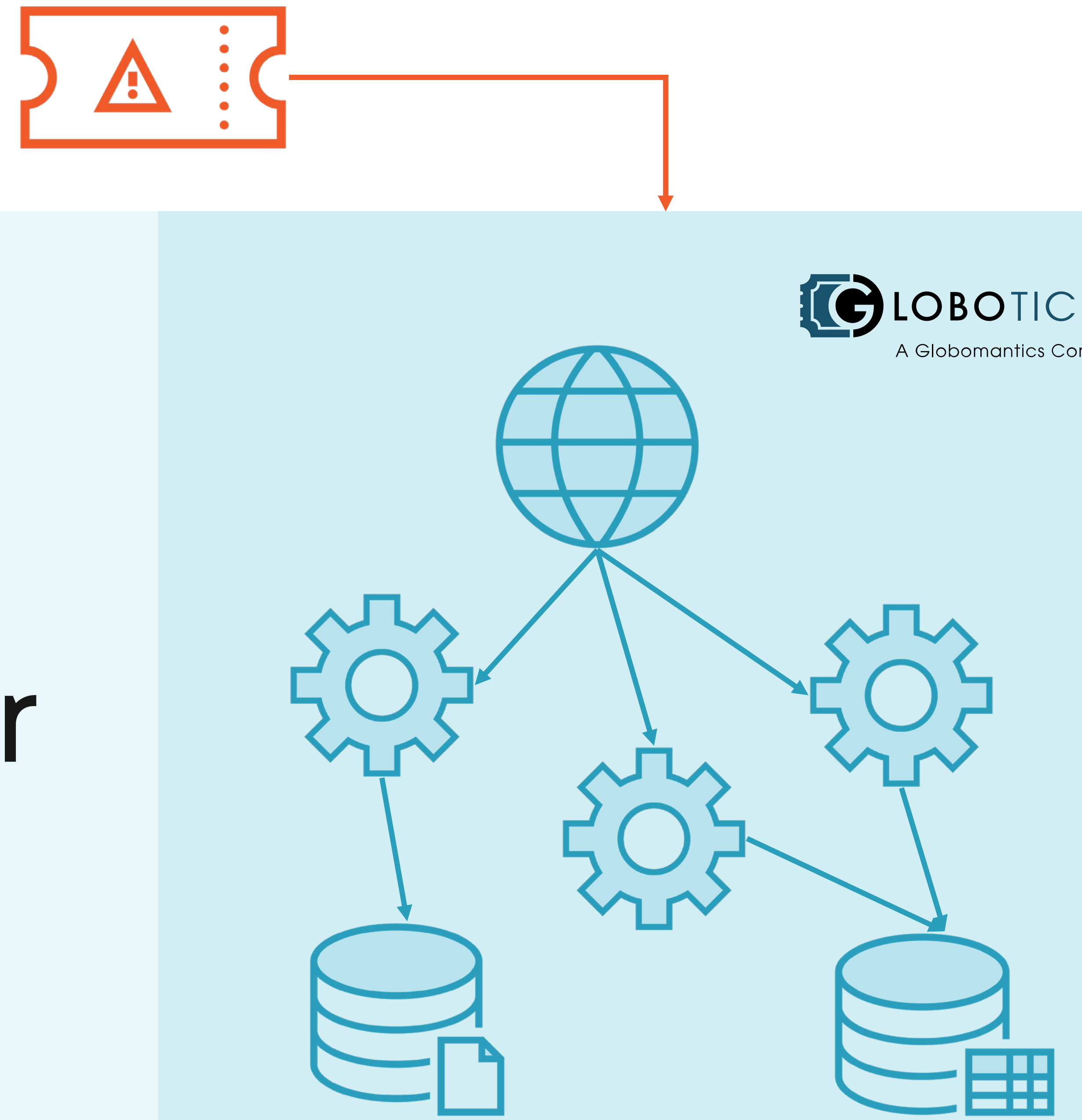




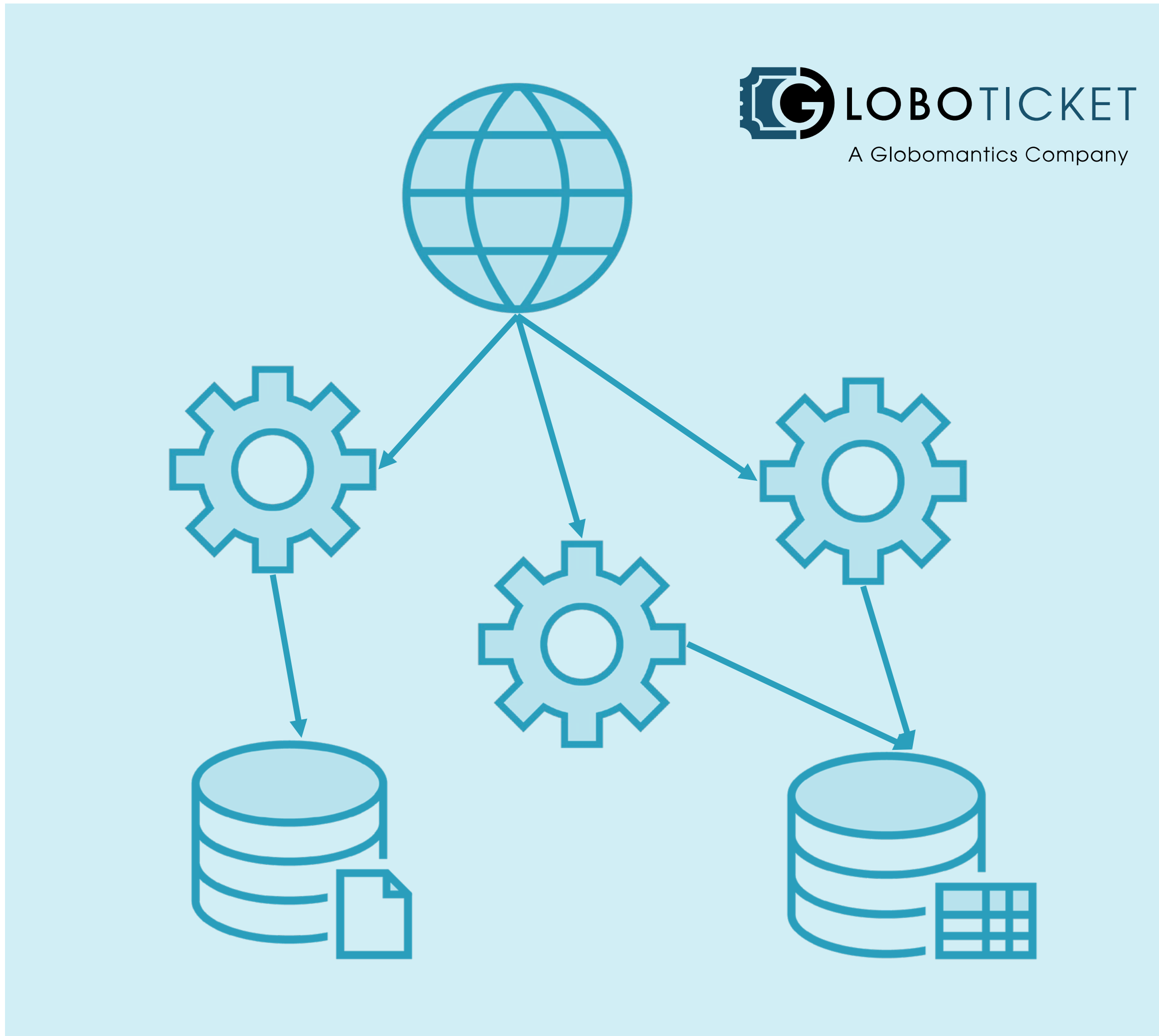
- Region
- Instance ID
- Version
- Customer ID
- Transaction ID

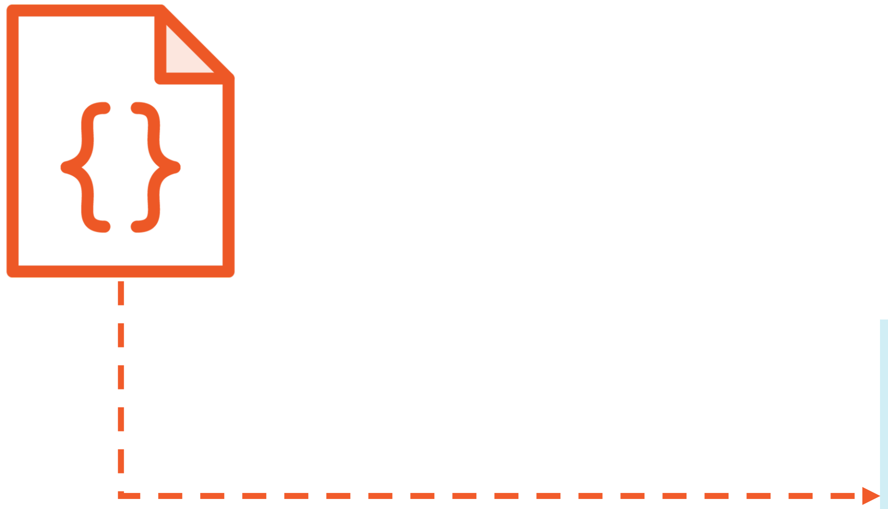
Scenario: Putting Logs to Use

How do you find the log for a unique error ID?

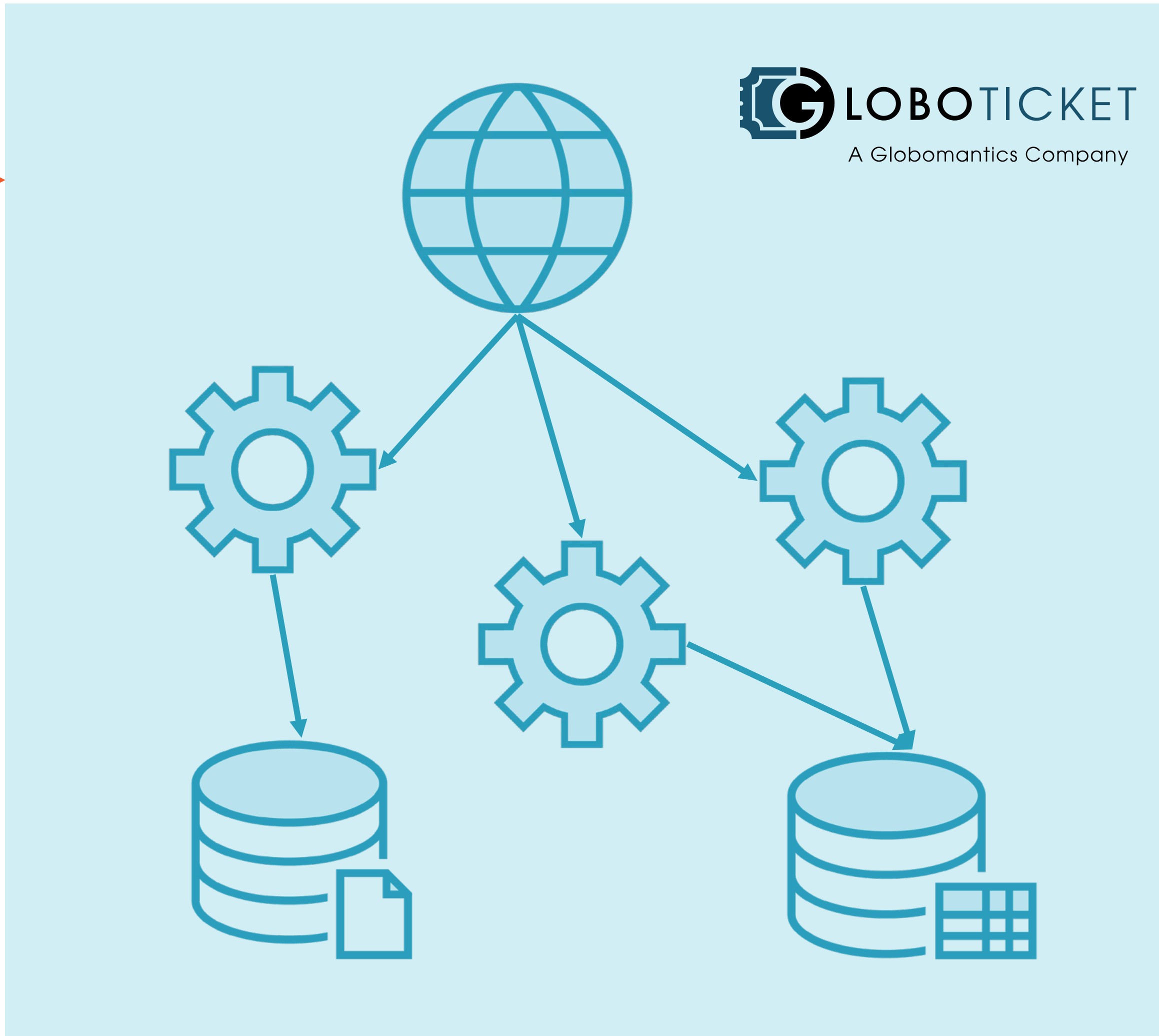


- All errors are logged
- UI for full-text search
- Component in metadata

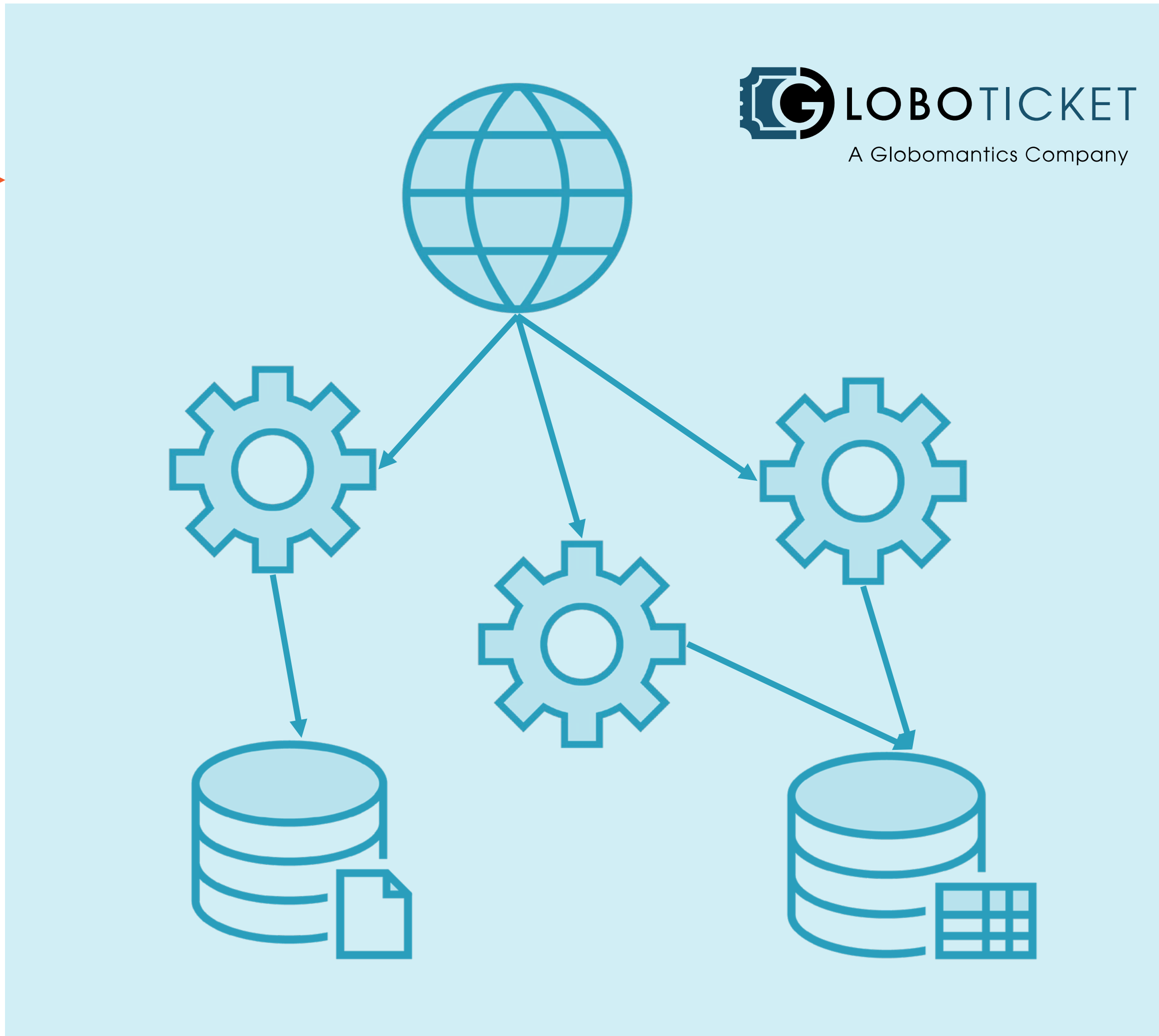




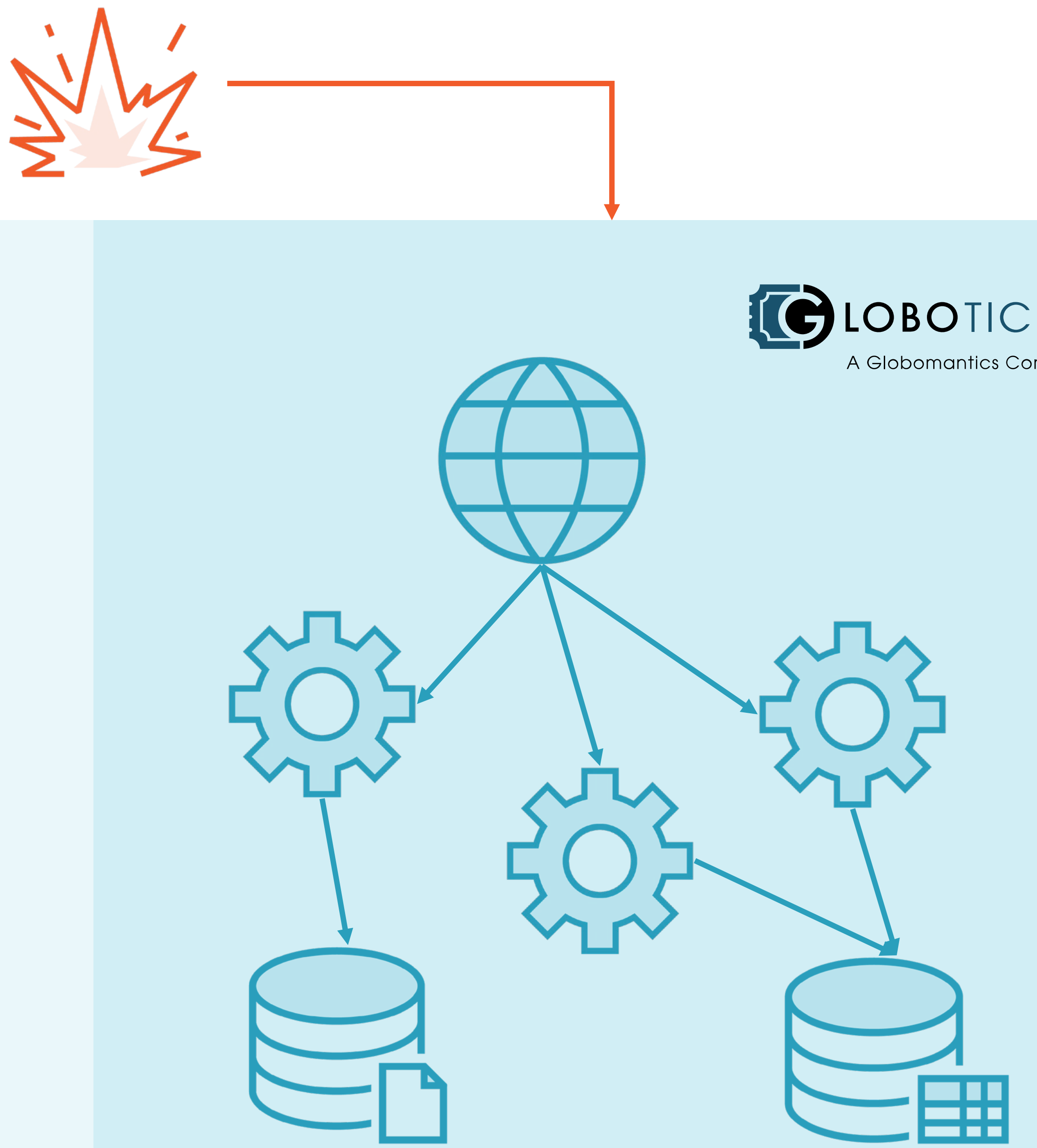
- Structured logs
- Key data in fields
- Efficient filtering

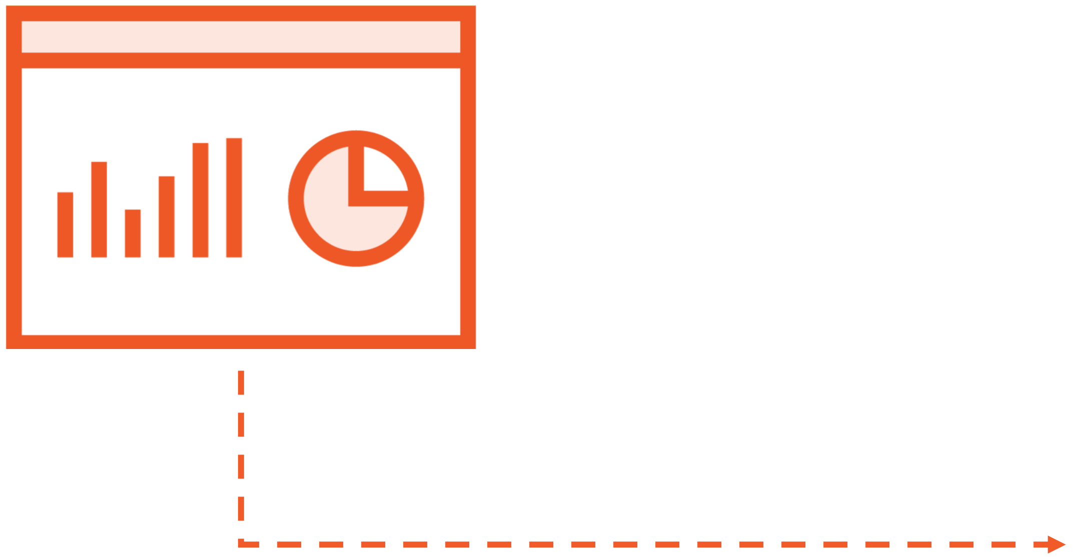


- More investigation
- Transaction ID
- Rebuild user path

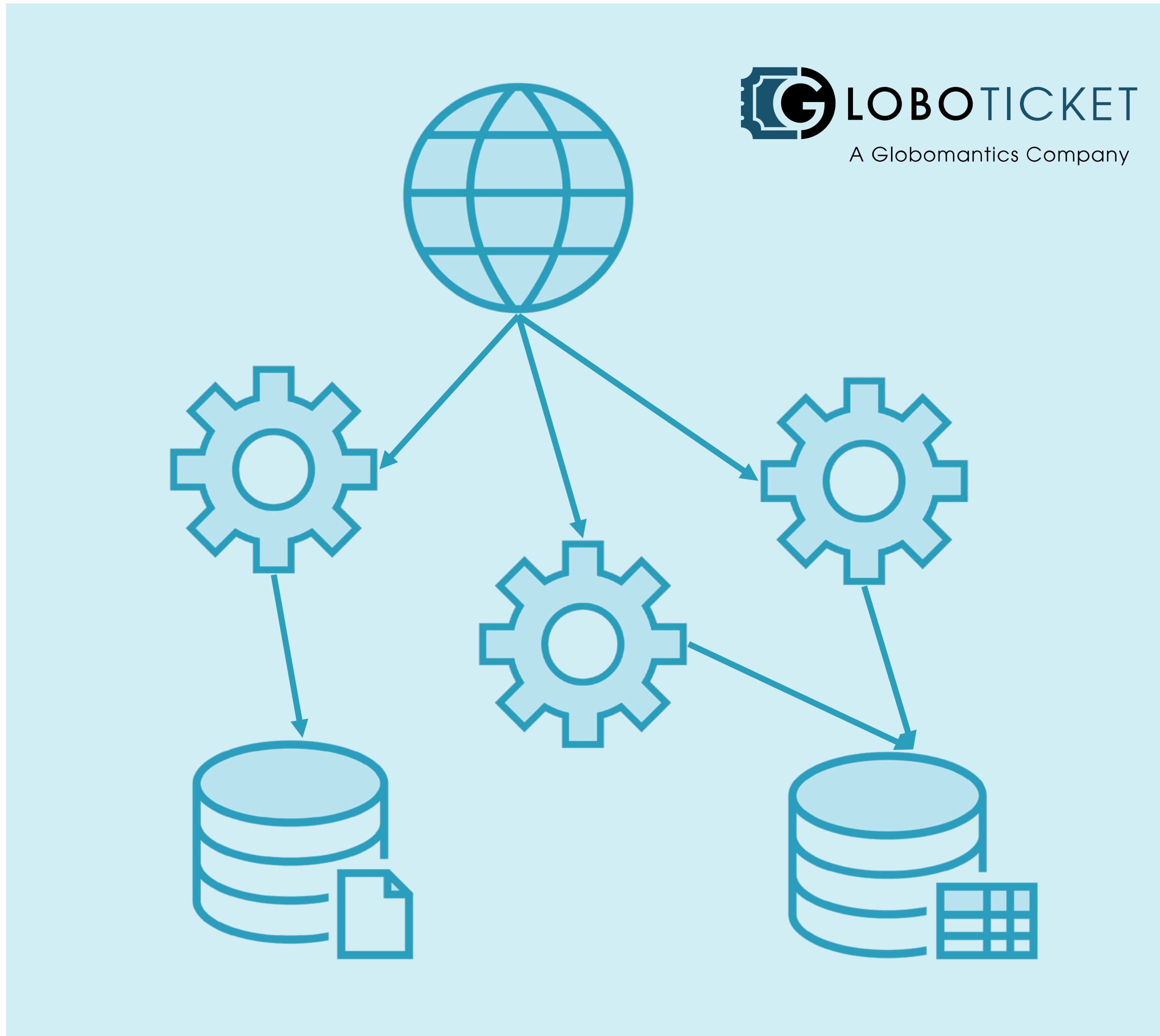


How do you track an increase in errors?

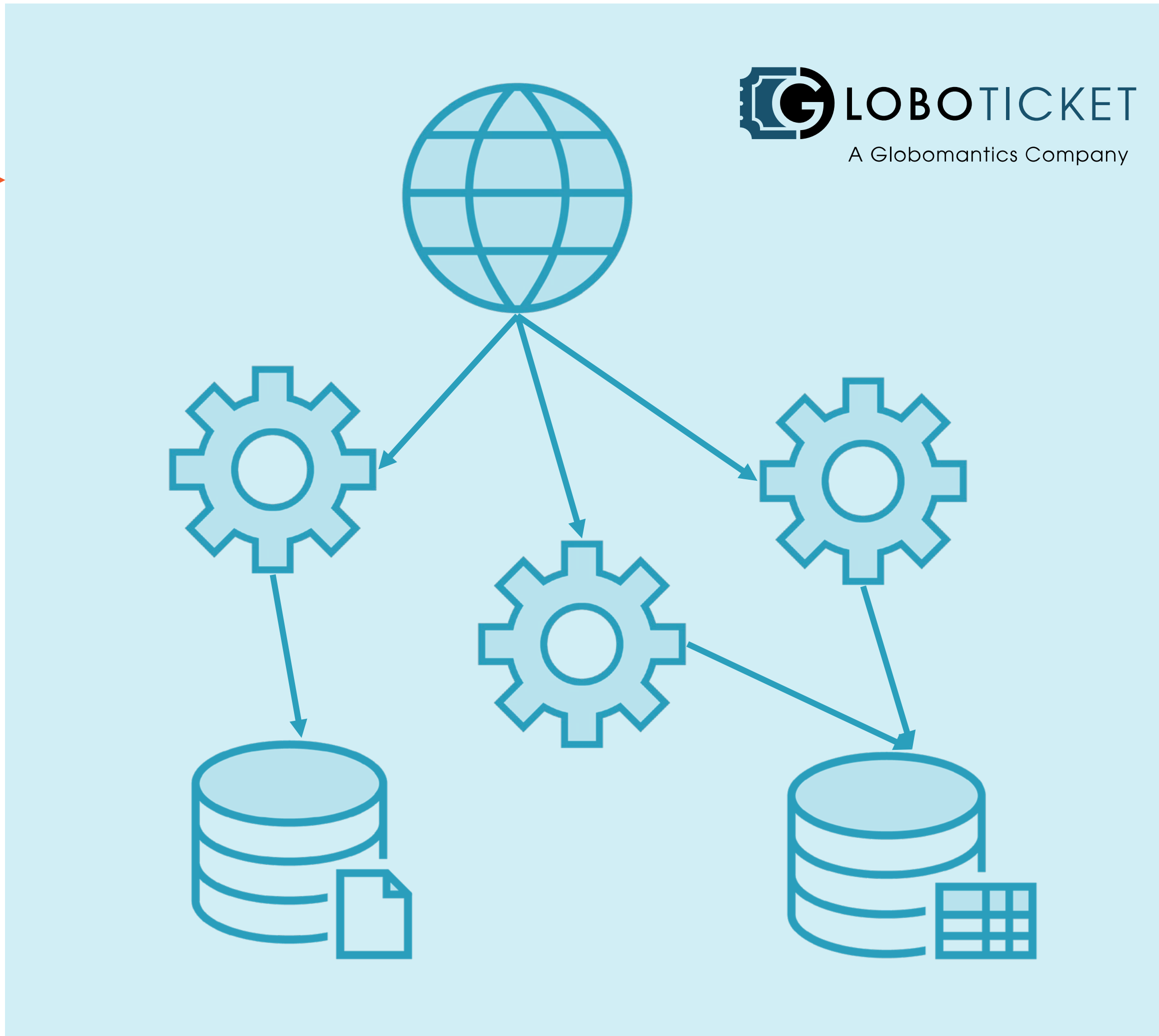




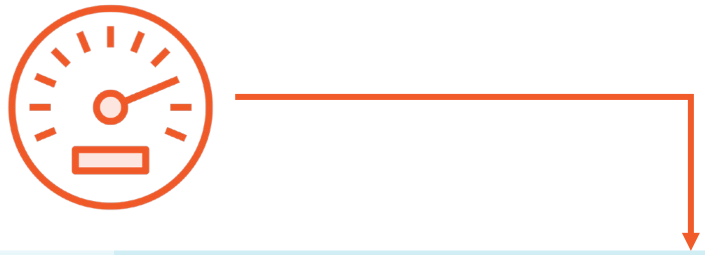
- Standard log levels
- Logging dashboards
- Alerts from monitoring

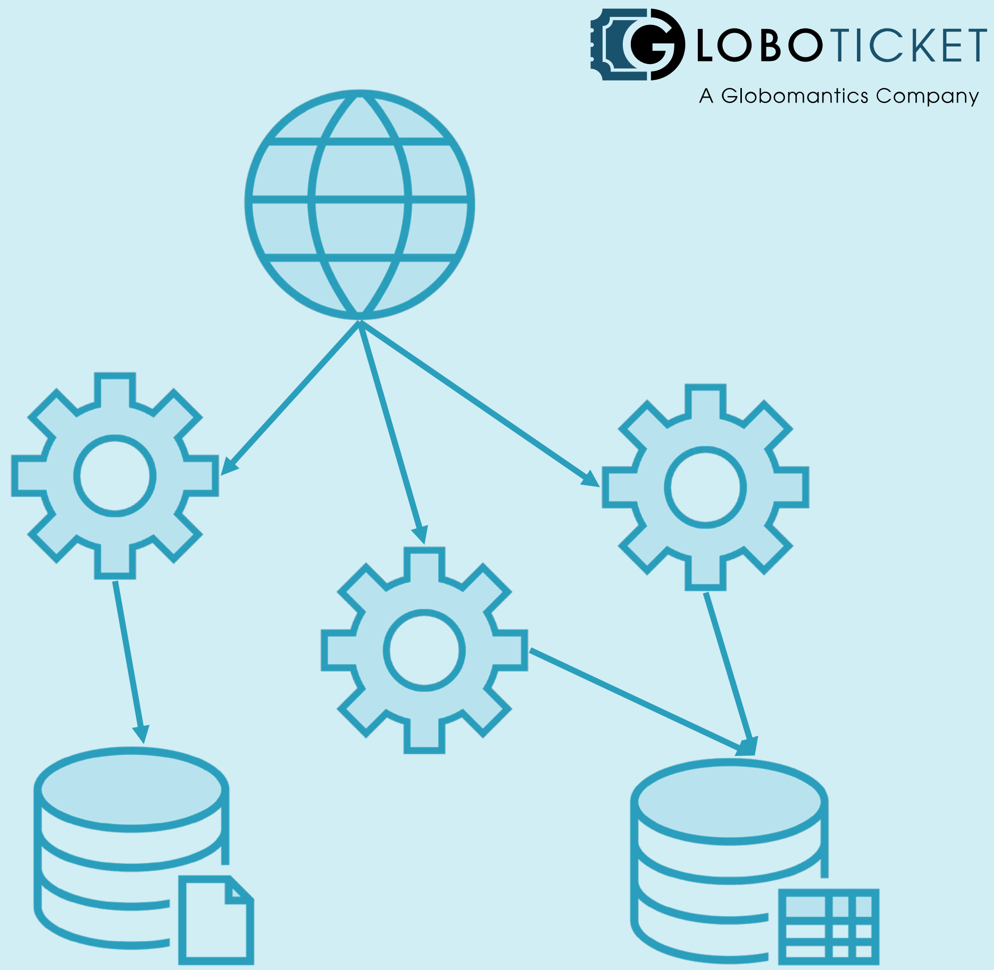


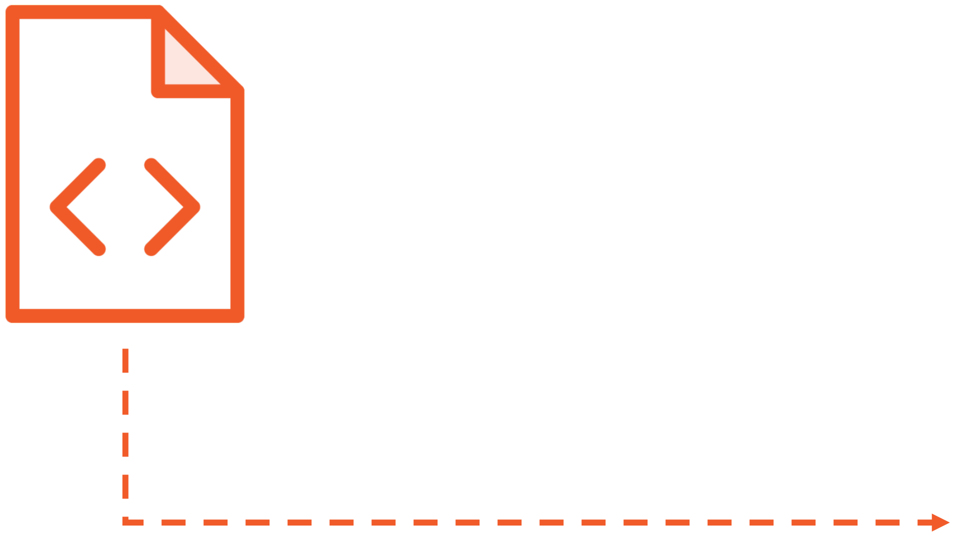
- Metrics for logging
- Count by level
- Trigger alerts



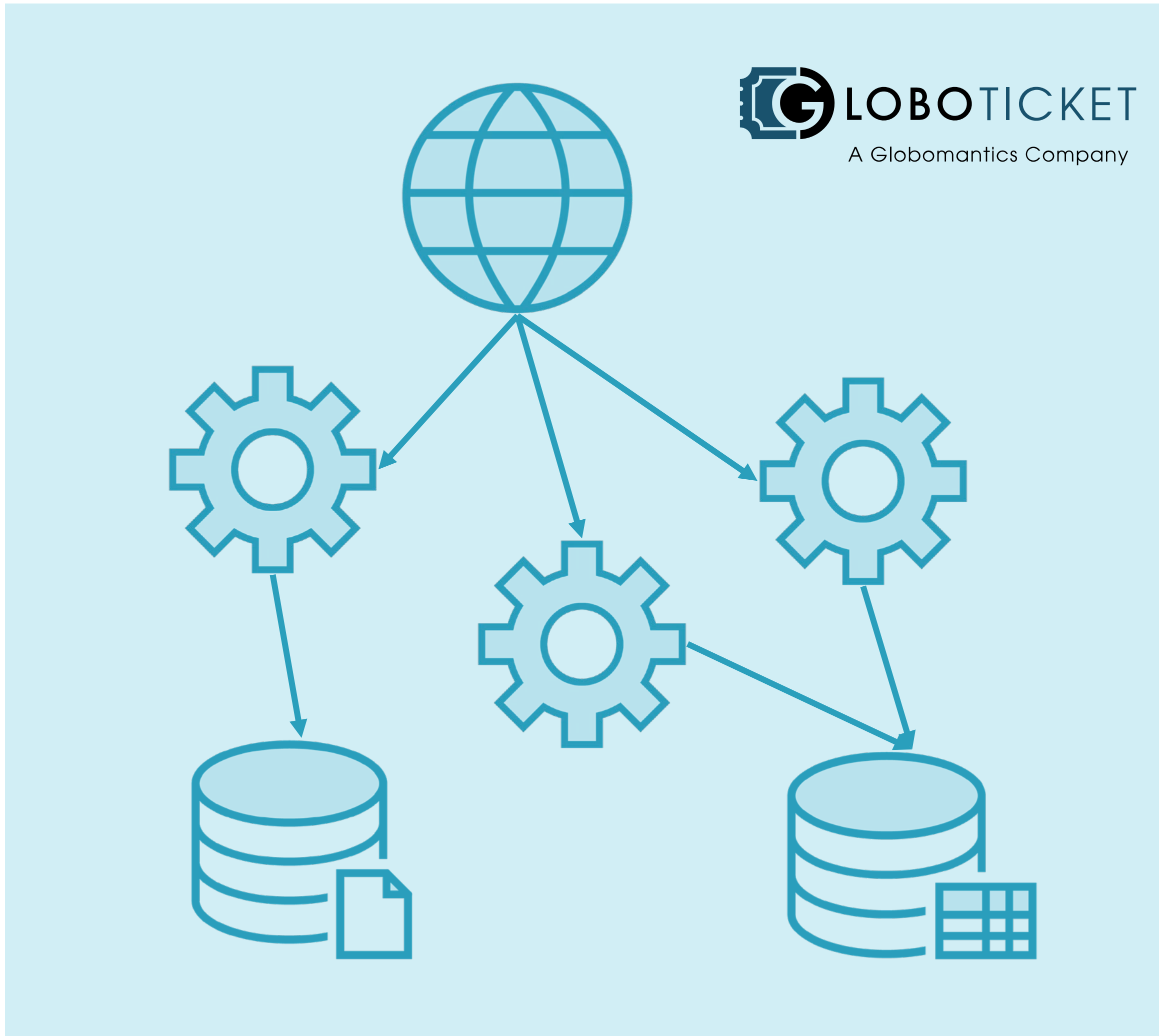
Can you increase logging levels on the fly?

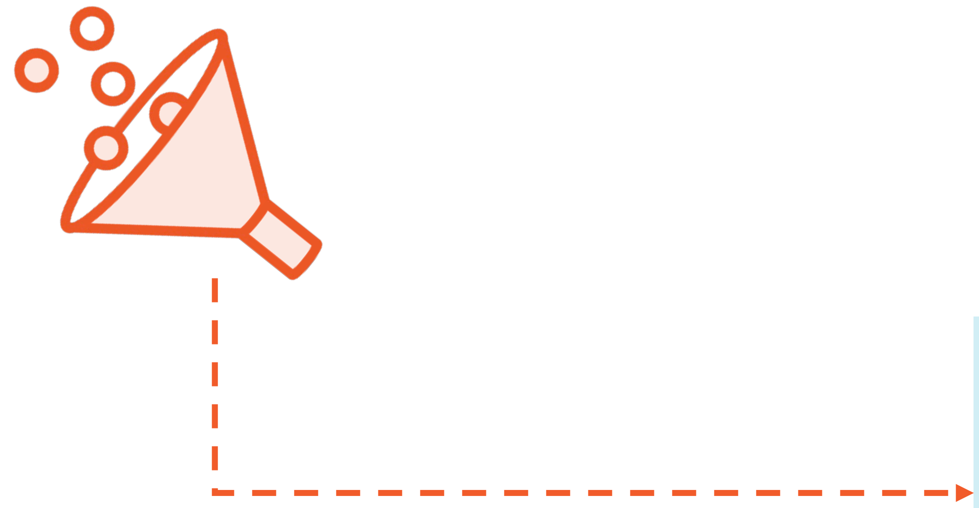




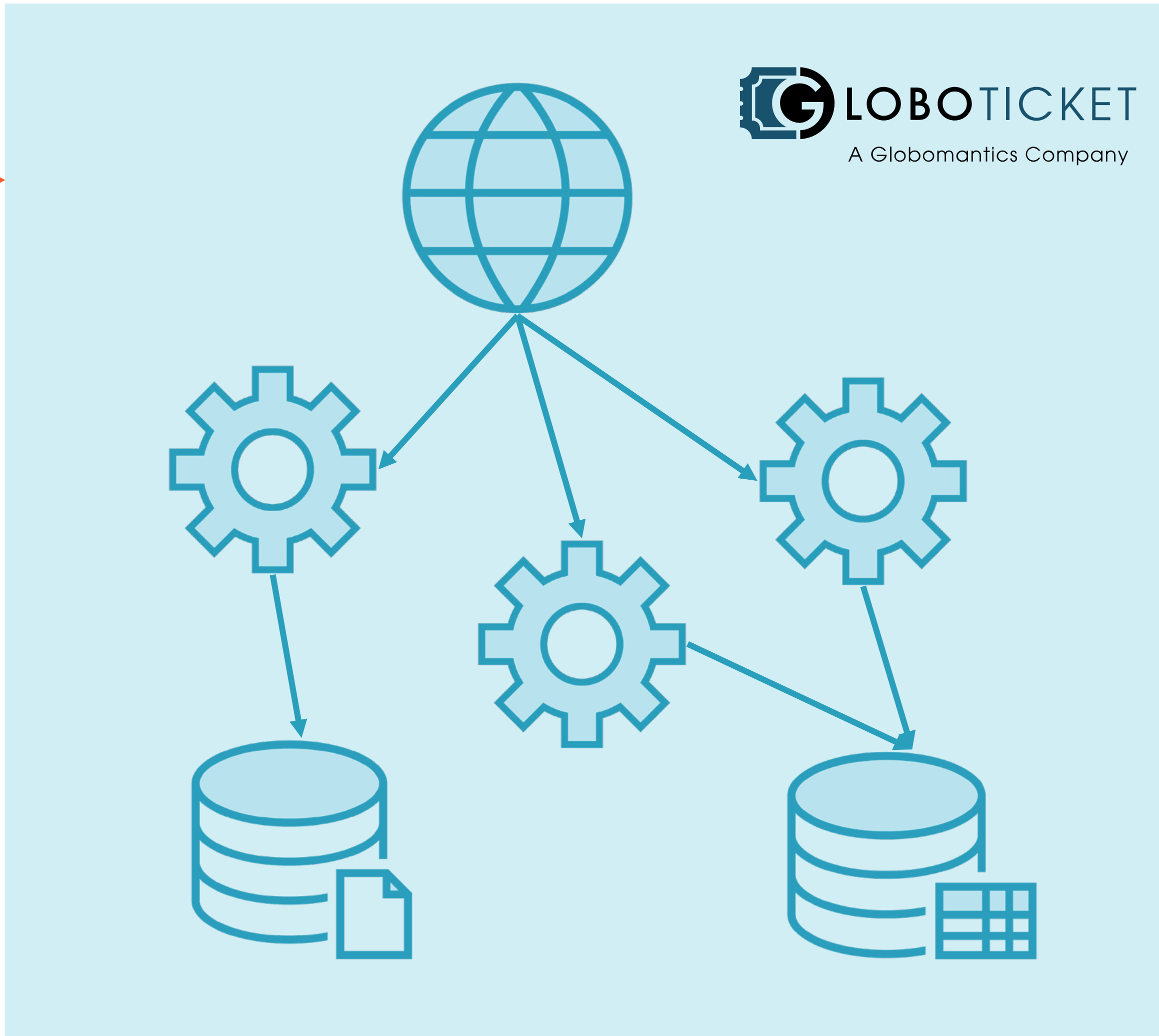


- Levels set by component
- Config rollout
- Instance restart

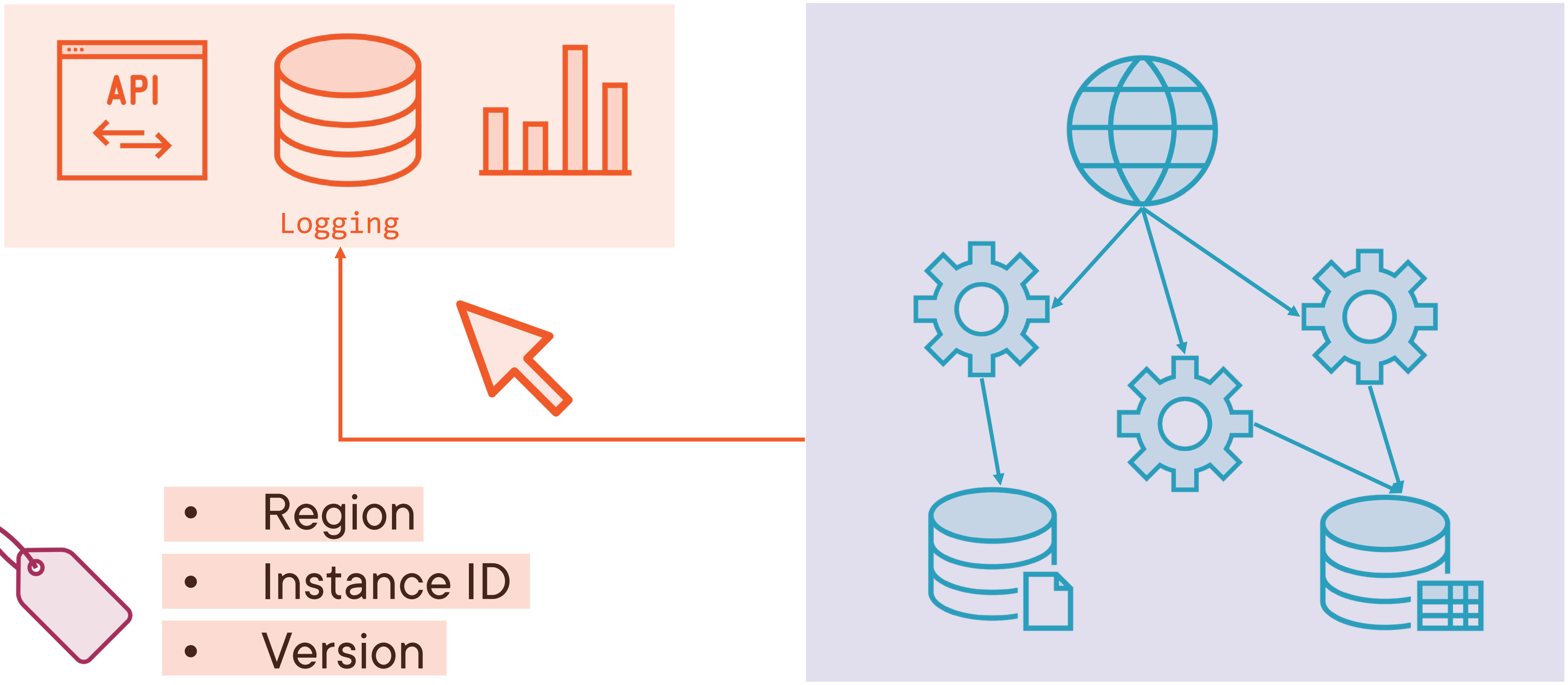




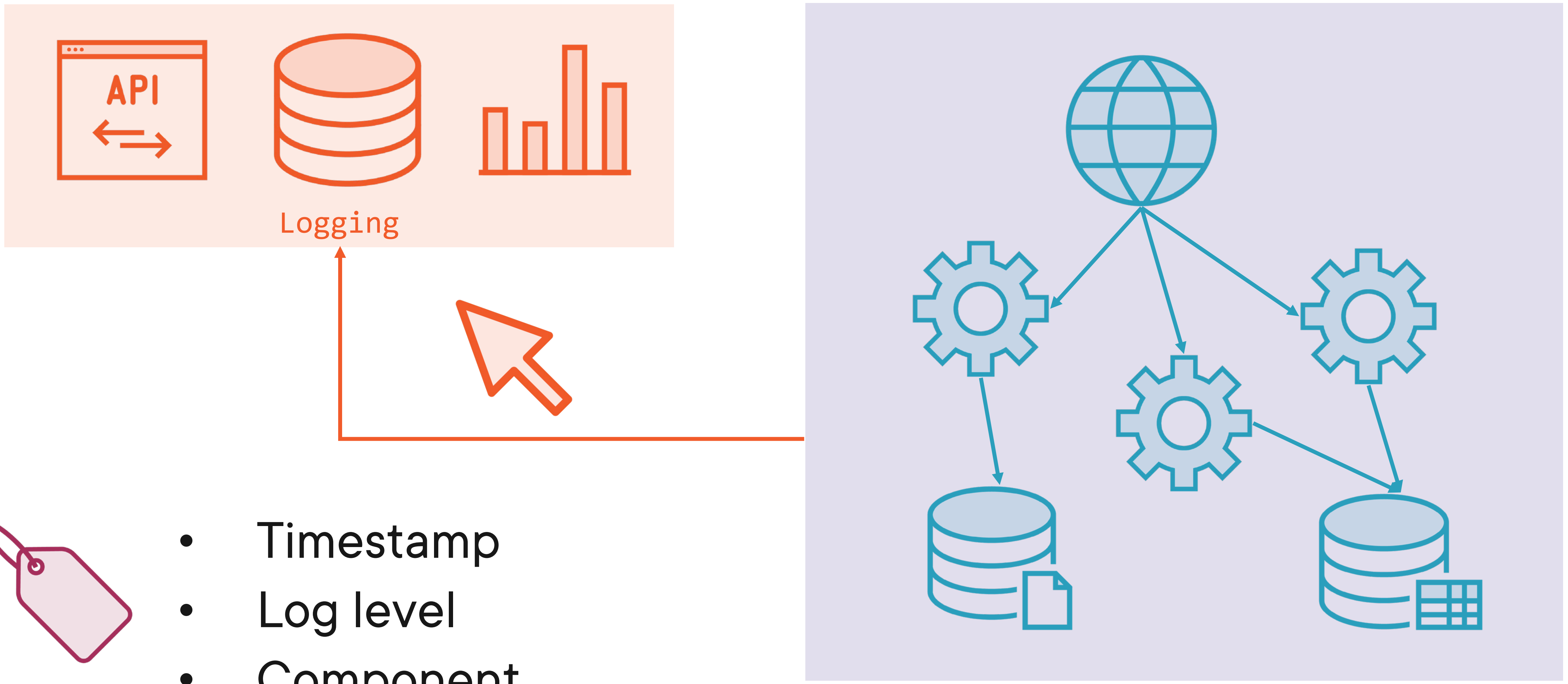
- Generate INFO logs
- Filter in pipeline
- Configure levels

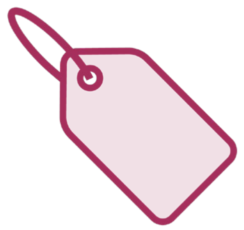


Module Summary



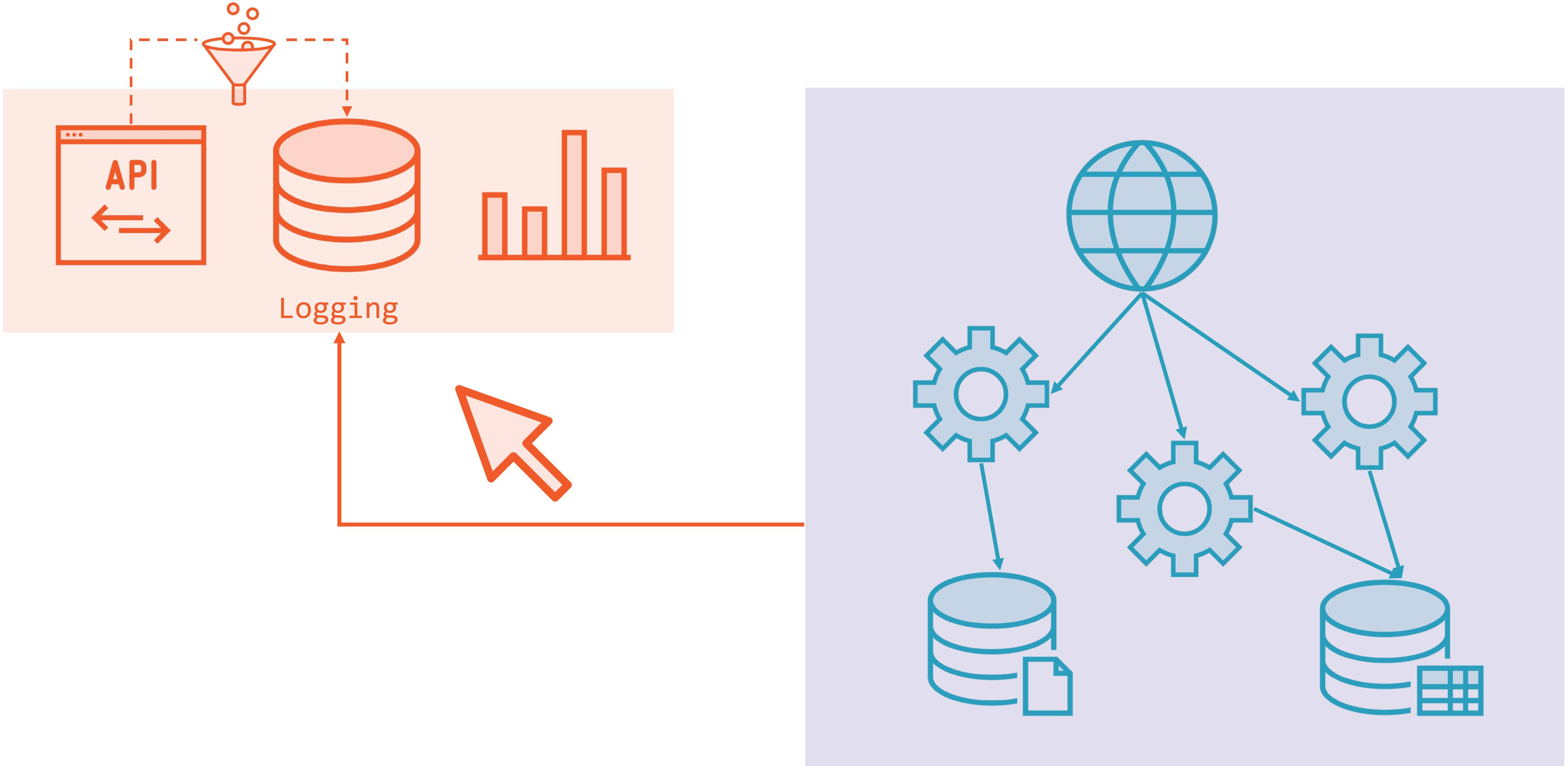
- Region
- Instance ID
- Version
- Customer ID
- Transaction ID





- Timestamp
- Log level
- Component
- Log entry

```
component=WEB & level >= WARN
component=API & level >= INFO
```

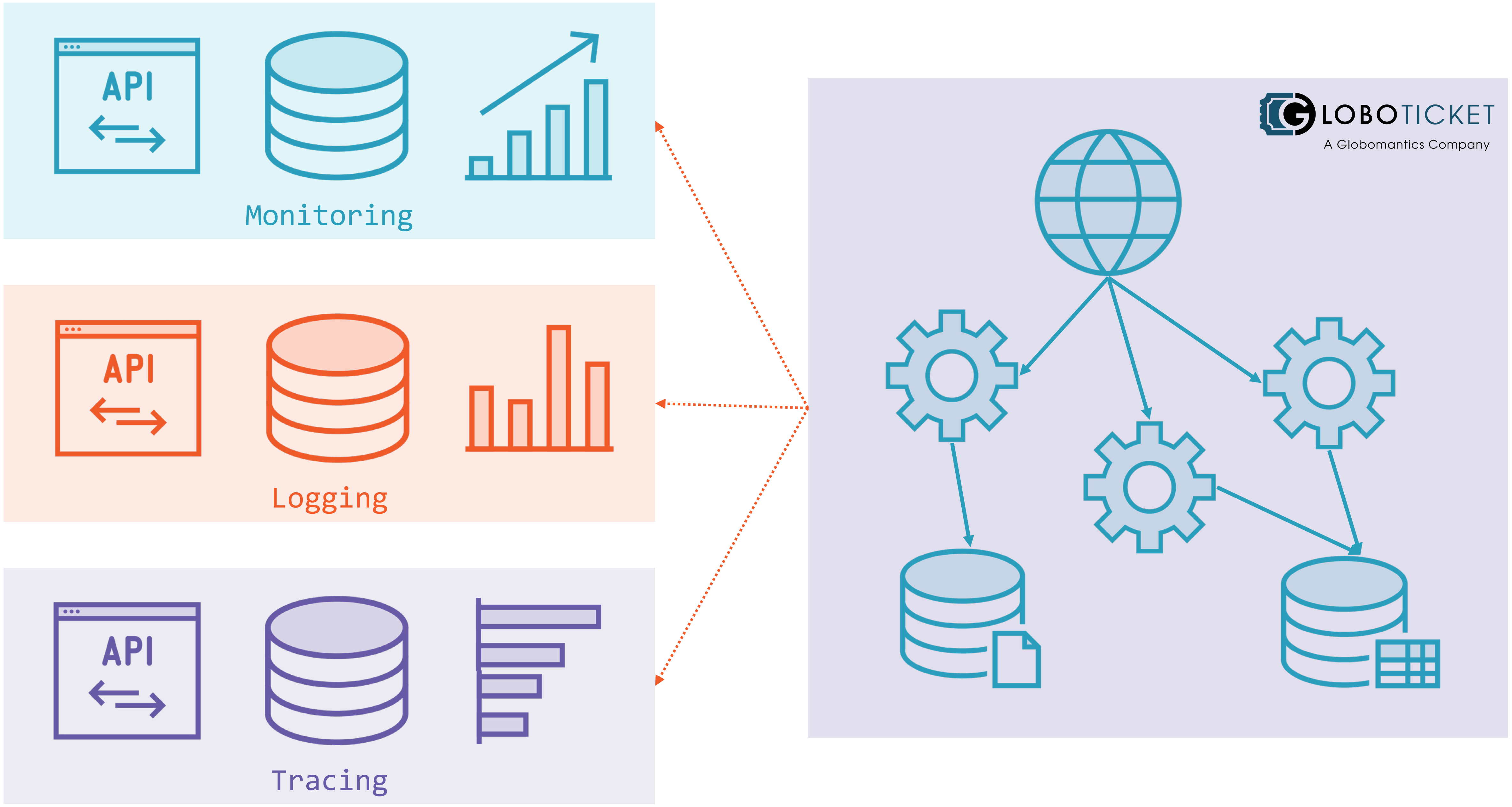


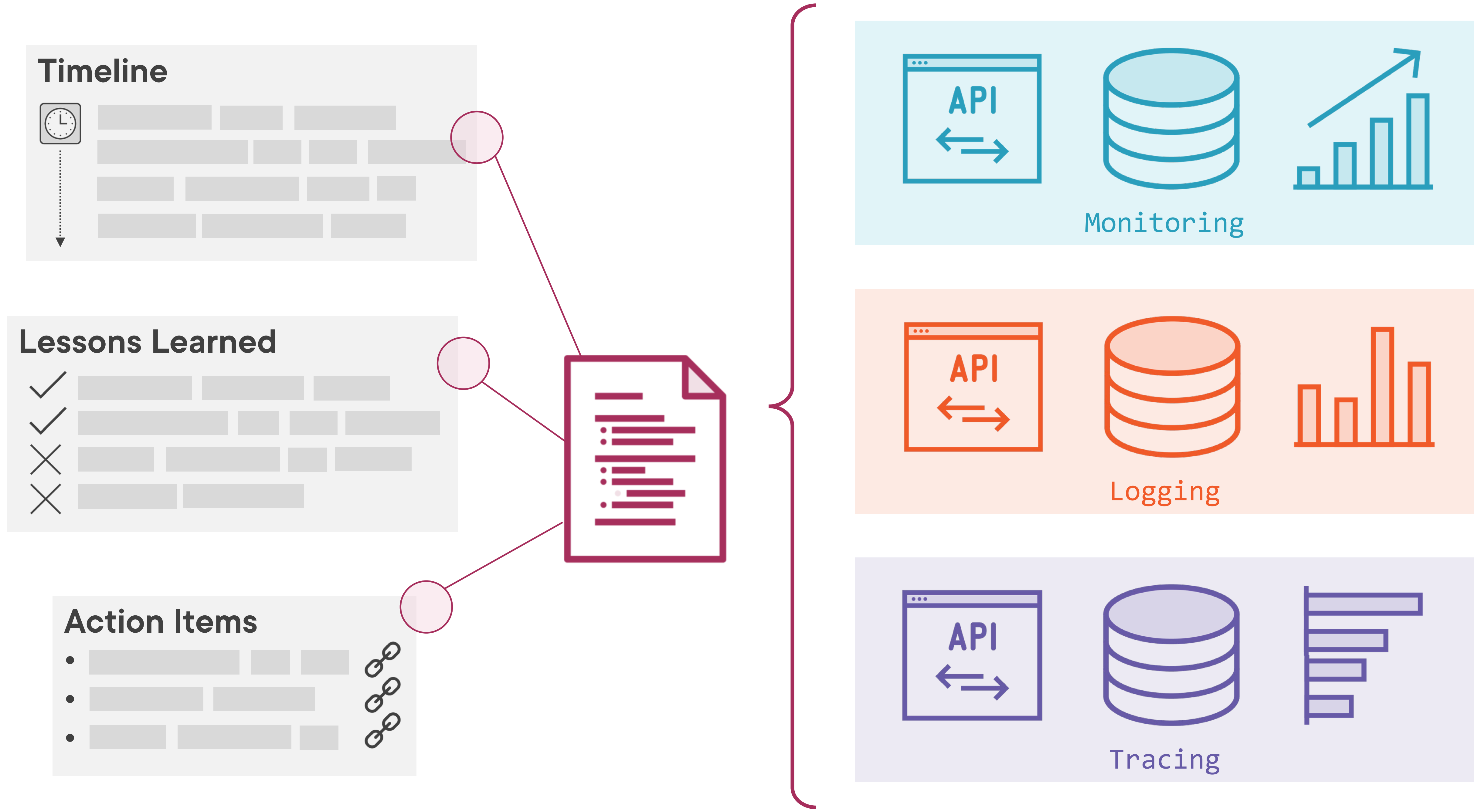
Designing Observability

Correlation

Real-time insight

Historical data


API
Monitoring
API
Logging
API
Tracing
GLOBOTICKET
A Globomantics Company


Timeline
Lessons Learned
Action Items
API
Monitoring
API
Logging
API
Tracing

Up Next: Driving Continuous Improvement with Service Levels