Contents

Scenario: Incident Management for Production Systems

Describe your process for handling critical incidents, including degraded modes, SLA breaches, and critical bugs in production.

Purpose

- Tests your ability to respond to high-severity incidents and maintain core functionality
- Evaluates your communication, troubleshooting, and stakeholder management skills
- Assesses your approach to root cause analysis, remediation, and prevention

---

Summary

Activate incident response. Evaluate impact and blast radius. Communicate transparently with stakeholders. Rollback or mitigate as needed. Provide degraded service if possible. Conduct root cause analysis and document learnings. Implement automation and process improvements to prevent recurrence.

Incident Detection & Triggers

- **Monitoring & Alerts:** Detect issues via real-time monitoring, alerting, and user reports
- **Triggers:**
  - System outages or high load (e.g., news feed degraded mode)
  - SLA breaches (e.g., uptime or latency below target)
  - Critical bugs impacting user experience, data, or security
  - Dependency failures or manual activation by operations

Response & Mitigation Strategies

- **Triage & Impact Assessment:**
  - Confirm incident severity and affected systems/users

  - Escalate immediately if user data, security, or revenue is at risk
- **Immediate Actions:**
  - Rollback recent deployments or config changes if needed
  - Apply hotfixes or configuration changes
  - Disable non-essential features (feature flags) to limit impact
  - Serve cached or simplified data (e.g., news feed snapshots, default ordering)
- **Degraded Mode:**
  - Skip non-critical processing (e.g., ML ranking, analytics)
  - Present default or cached content to users
  - Ensure UI remains functional and clear

Communication

- **Internal:**
  - Notify engineering, product, and leadership teams immediately
  - Provide regular updates in a shared incident channel (e.g., Slack, incident dashboard)
- **External:**
  - Update status pages and communicate with affected customers if required
  - Be transparent about impact, mitigation steps, and estimated time to resolution
  - In-app messaging to inform users of degraded mode or limited functionality

---

Root Cause Analysis (RCA)

- **Data Collection:** Gather logs, metrics, and relevant artifacts
- **Analysis:** Identify technical and process root causes; involve relevant teams
- **Reproduction:** Attempt to reproduce the issue in a controlled environment

---

Postmortem & Prevention

- **Blameless Postmortem:** Document what happened, why, and how to prevent recurrence
- **Action Items:**
  - Implement automation, new tests, or monitoring to catch similar issues early
  - Update runbooks and incident response procedures
  - Share findings and action items with the broader team

Metrics for Success

- Time to detect and resolve the incident
- Number of users affected
- Communication response time
- Recurrence rate of similar incidents
- User engagement and retention during incidents
- Time to full recovery

Trade-offs & Limitations

- Reduced personalization or feature set during degraded mode
- Lower engagement or satisfaction during incidents
- User trust depends on transparency and speed of recovery