Riphah International University

Site Reliability of Engineering (SRE)

Assignment no 1

Submitted by: Muhammad Qasim

Sap ID: 37137

Section: BSCS-6A

Submitted To: Javaria

Riphah School of Computing & Innovation Faculty of Computing Riphah International University, Lahore 2023

Case Study:

On Tuesday, February 28th, 2017, the US-East-1 region of Amazon Web Services S3 saw a complete outage from 9:40 am to 12:36 pm PST. AWS S3 (Simple Storage Service) is a cloud object storage solution that many services rely on to store and retrieve files from anywhere on the web. In addition, many other AWS services that depend on S3 — Elastic Load Balancers, Redshift data warehouse, Relational Database Service, and others — also had limited to no functionality. Similar to the repercussions of the AWS outage caused by a route leak in 2015, the S3 outage disrupted a large number of services that depend on AWS over roughly 3 hours. These services included Quora, Coursera, Docker, Medium, and Down Detector.

1. **What was the underlying cause of the outage?**

Ans:

February 28, 2017, the AWS S3 outage was caused by a human error that happened during the debugging process. When addressing a billing system performance issue, an AWS team member inadvertently deleted servers essential to S3's functionality. When a command was mistyped, it took down a bigger set of servers instead of just a few. This mistake caused an S3 subsystem to malfunction. S3's reliance on distributed systems meant that it could not recover from the loss of a significant server section, which resulted in a cascade effect and extended downtime in the US-East-1 region.

2. **How was the issue resolved by AWS?**

Ans:

An organized strategy was used to fix the AWS S3 outage:

Root Cause Identification: During debugging, AWS identified a human mistake as the root cause.

Restoration Efforts: The engineers concentrated on restoring the impacted S3 subsystem, trying to recover servers that had been unintentionally deleted.

Mitigation Strategies: To reduce disruption and offer updates, AWS put in place failover measures and kept in touch with customers.

Post-Outage investigation: To identify weaknesses and stop reoccurring events, a thorough postmortem investigation was carried out.

3. **What proactive measures could have been taken to avert such an incident?**

Ans:

Better Testing:

Make sure that tests for modifications to important systems are done better.

Automation and Safety Measures:

To minimize human error, increase automation and put in place safety measures.

Change Management:

Documentation and peer evaluations help to fortify change management procedures. Enhanced redundancy and failover capacities are necessary.

Monitoring and Alerting:

To spot abnormalities in real time, improve your monitoring and alerting systems.

4. **Reflect on the key takeaways derived from this incident**

Ans:

The AWS S3 outage emphasizes how critical it is to combat human error by implementing strong security measures and comprehensive testing protocols. Automation is essential for lowering risk and increasing productivity. Peer reviews and other change management procedures are crucial for guaranteeing that changes are implemented correctly. System resilience depends on both redundant service and failover capabilities as well as efficient monitoring and alerting platforms for prompt response. Ongoing instruction and training improve team dynamics, and chaos engineering techniques proactively spot flaws. Prompt incident resolution is facilitated by open communication and openness. Postmortem analysis promotes ongoing development and aids in averting recurrence.