smartQED

Fix it Fast

5 Key Challenges in Handling IT Incidents & How smartQED Accelerates Resolution

A smartQED Whitepaper

Julie Basu, PhD

CEO & Head of R&D

Table of Contents

1. Incidents in IT Services and Their Impact

In today's technology-dependent world, rapid resolution of IT incidents is more critical than ever before.

Many enterprises depend heavily on IT Services, both for their internal operations as well for external customers and communications.   For example, if the order entry service is down for an e-commerce site, it affects the revenue and/or reputation of the enterprise adversely.

IT service providers have the additional responsibility of adhering to SLA agreements with customers, where strict time limits might be in place for resolution of IT incidents. Violation of such SLA agreements can result in monetary penalties for the provider, as well as loss of goodwill with their customers.

War Rooms and Fire Fights

Accordingly, speed of problem solving is of paramount importance for an IT incident. Time truly translates to money in most of these cases, and typically, the head of IT operations starts a 'war room' or 'fire fight' to investigate the issue on an urgent footing.

Compounding the time pressure is the fact that IT system architectures have gotten quite complex in recent years. They have evolved to handle millions of online users, and thousands of transactions per hour with high reliability and 24x7 availability. A typical web application can involve multiple services or microservices that are deployed on a single or hybrid cloud, and it can use load balancers for scalability. Development teams often follow agile methodologies, causing different parts of the system to be changed independently and frequently.

For example, let's consider an e-commerce site that sells products. Their application talks to an on-premise product inventory database, has an order entry service that tracks customers and their orders in a cloud database, and integrates with an external SaaS payment service as well. Application code for the order service may have been upgraded recently, and code for the SaaS payment service may also have been changed at around the same time.

When a problem occurs in such a complex system with many moving parts, it is difficult to identify which exact component or activity is responsible for the issue. Therefore, subject matter experts (SMEs) for the various components need to be pulled into the investigation, and often their managers as well.

Based on what we have heard from our customers, and from our own deep experience in developing and supporting Enterprise IT systems, the size of such investigation teams depends on the complexity and severity of the problem at hand and the company size. It can range from a dozen admins/engineers & supervisors all the way to several hundred people (about 1100 is the maximum we have heard). Each team member is a stakeholder in one way or another, and they are all interested in following the progress and eventual resolution of the incident.

Not unexpectedly, communicating and collaborating with such a large group of people under high time pressure raises significant challenges. Understanding these issues is the main subject of this whitepaper. But first, let us see some real-world examples of major outages that have occurred recently and made it to the news headlines.

Recent Examples of Major IT Incidents

Costco faced a major outage during Black Friday sales in November 2019, with their online shopping site giving error messages that customer orders would be delayed. As reported in the news (see reference [1]), the outage lasted a whopping 16+ hours! Thousands of customers were affected by it and it resulted in the company losing millions of dollars.

Another high-profile outage was reported for Amazon during its Prime Day 2018. Customers were unable to complete their orders due to a surge in website traffic. News reports mentioned over 300 people were pulled into an emergency conference call and it was 'chaotic' (see reference [2]). People were 'scrambling' to solve the issue, which took hours to do.

SORRY
something went wrong
on our end

Please go back and try again
or go to Amazon's home page.

RoRo
Meet the dogs of Amazon

Such major incidents are in fact quite common in IT, and very few are the subject of news articles. Even a non-major incident may have significant impact in terms of the time and resources it consumes. Every minute counts in the investigation process, as we shall see next.

The Cost

The foremost question in the minds of executives when a major incident happens is: “How much is this problem going to cost us?”

The cost of an incident involves multiple factors, including:

(1) Business revenue lost
(2) Cost of resources to investigate the problem
(3) Loss of goodwill and brand reputation

The first two factors are one-time considerations, whereas brand damage typically has longer term impact – it may have semi-permanent or even permanent effect in some cases. Additional costs of the incident include loss of productivity and/or revenue for impacted users.

Business Revenue Lost:
How much revenue is lost depends on the nature of the business and the type of IT application. A study from the Ponemon Institute (see reference [3]) considered 63 data centers and found that on average, an unplanned outage costs nearly **$9,000 *per minute***. This cost is of course bigger for larger companies with higher number of users, and it was observed to be increasing from 2010 to 2016.

Cost of resources to investigate the problem: During the incident investigation IT engineers and managers are diverted from their regular tasks and activities to focus on the fire fight. Planned tasks are delayed by unexpected incidents, and some people may have to be paid overtime if the outage lasts beyond normal business hours. Long term / indirect costs include high stress on personnel and excessive turnover.

Loss of Goodwill and Brand Reputation: This cost generally cannot generally be quantified in terms of money. Losing customer goodwill can affect both current and future revenue. If a popular internet business or service is down, social media will definitely be involved. Impacted users can be expected to discuss and share experiences on various social media including twitter, online discussion forums and interest groups.

Social media presence and image are of prime importance for businesses today – they are essential for brand building. But negative comments on the internet are often there to stay – history cannot be easily erased!

2. Key Challenges in Handling IT Incidents

In this whitepaper we discuss several key challenges in resolving IT incidents, with specific focus on collaboration overheads and time delays from confusion and chaos during the investigation process. Let's get started!

Lack of Shared Understanding

Complex problem investigations require multiple teams with different domain expertise all working together to come to a shared understanding of what is going on and how they should try to fix it.

Fragmented Collaboration

Teams use various tools such as call bridges, chats, emails and issue tracking systems to collaborate and coordinate actions. However, such information can easily get fragmented, and manually searching it can take up a lot of precious time. As a result, people may not be on the same page and/or it may take them long to do so.

Difficulty in Understanding the Status

Additionally, all the stakeholders need to read the serial ticket and/or chat updates and build up a 'picture' of the investigation progress and current status in their minds. This is time-consuming and error-prone to say the least. Text-based updates, however detailed, generally do not provide enough clarity, because the information needs to be read and pieced together to see the whole picture. Accordingly, managers and executives may need to be updated manually over the phone or conference calls, which adds a significant communication overhead during the investigation process.

Larger team sizes make reaching shared understanding a bigger challenge, as collaborating effectively becomes more difficult. The more the merrier? Not always!

Too Much Text to Read

Most enterprise IT teams track incidents using tools such as ServiceNow or JIRA. A major difficulty in using these tools is that they are text-based and linear. An incident ticket for a complex problem can get very long and convoluted, which is a burden to read and understand under time pressure.

Additionally, for global teams not all members might be native speakers of English, and text-based information can get misinterpreted or 'lost in translation'. This can cause confusion, duplicated work during shift changes & handovers, and delays in getting the problem fixed.

Ad-Hoc And Unclear Investigation Strategies

Once a major incident / problem occurs, the investigation process involves humans who need to figure out what is going on and how they should try to fix it. As discussed, enterprise IT problems can often be complex and cross-functional, and investigations may involve many SMEs who are working remotely from different geographic locations across various time zones.

Identifying the cause of a problem generally is a process of elimination, much like a murder mystery – health states and symptoms of components likely to have caused the issue are examined by investigators and they serve as evidence to clear or confirm whether a component is responsible for causing the problem.

All-important questions for supervisors and incident managers during the investigation process include: "Are we following a systematic investigation strategy? Is the right team involved? Who should we call next?"

Where do we start?

Unless the symptoms point very clearly to one or more specific components as causes, the starting point for the investigation can be unclear. Depending on the complexity of the IT system, many teams may need to be involved to check out their respective parts of the system.

Unfortunately, war rooms are commonly known to encounter the difficulties of 'blame storming' and finger pointing. In fact, problems can often be bounced back and forth among teams with obvious reluctance to take responsibility – human psychology is such that we tend to avoid taking the blame for causing a problem.

What's Next?

Assuming the problem is complex, it is essential to have a systematic investigation strategy that is comprehensive and is clear to all. This strategic knowledge typically resides in the heads of admins and domain experts, and is often not well-communicated to everyone in the team.

As a result, people may be performing irrelevant work that either does not help or results in unintended delays. They may not even be aware of necessary work aligned with the overall strategy that would help resolve the incident faster.

Inadequate Knowledge Sharing & Inefficient Reuse

This is a challenge that most managers grapple with – "How do we ensure that the lessons learned are incorporated in the best practices and utilized effectively for future problems?"

Problem-Solving Expertise is Often Siloed

How to approach a problem quickly and effectively is knowledge that often resides in the heads of specific domain experts. It is built up from their years of problem-solving experience, much like a senior medical doctor. However, sharing such knowledge may not be a priority for them, with the result that new members in IT teams are not adequately trained up. This lack of knowledge-sharing and training may cause significant delays when a crisis happens in the IT service, especially when the experts are not available.

Knowledge Articles Are Inefficient

The traditional way of capturing problem-solving experience is to write knowledge articles and trouble-shooting guides. These articles can be rather complex, explaining the various circumstances, and prescribing solutions for future problems. While the intent is good, the trouble with this approach is that others have to first read and interpret these articles properly, then apply them appropriately at the right time. Needless to say, this is not an easy task, especially for people that are new to the team.

Terry Gallagher, a seasoned Crisis Manager for IT incidents at large enterprises, a co-founder of smartQED and the head of Customer Solutions, says that in his experience, although much effort is put into writing them, more than 80% of an organization's knowledge articles are never opened, much less used.

Shortage of Skilled Engineers & Admins

New technologies are emerging today at an ever-increasing pace, including many different types of cloud-based infrastructures and services. It is difficult for IT teams to keep up with the latest technologies and master them in a short time.

Adding to this challenge is the fact that senior IT engineers and admins are much in demand and they may choose to change jobs at any time. This fact alone can keep IT managers up at night, worrying about how to handle such a situation should it arise. When they leave, seniors take most of their problem-solving knowledge with them, as there are no good ways today to capture and communicate it effectively to the team.

3. How smartQED Accelerates Incident Resolution

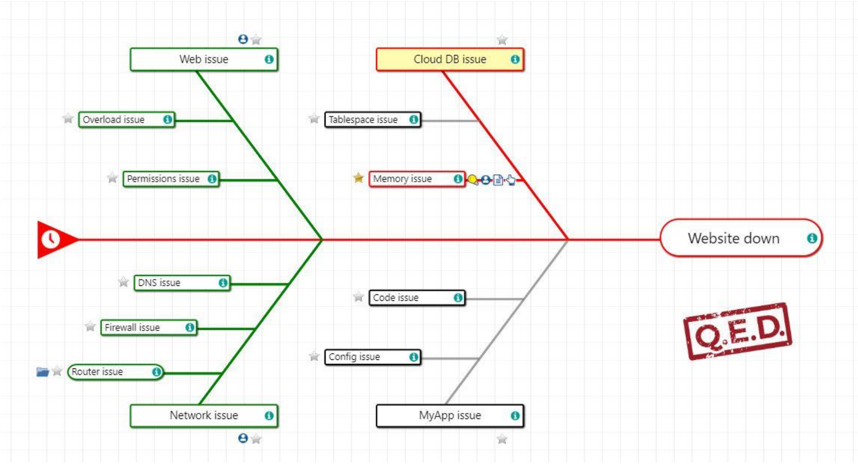
Our mission at smartQED is to solve the challenges and delays in IT incident resolution effectively. Based on our in-depth industry experience, we believe that new approaches are needed. In smartQED we have introduced specialized visual Investigation Maps™ for efficient team collaboration while solving problems jointly, and the augmenting of human intelligence with powerful machine learning to help accelerate incident resolution with recommendations from historical data. Our innovative approaches go way beyond just keyword search or text analytics on incident tickets.

Collaborative Investigation Maps™

One of our key insights is that reading and writing text in incident tickets to communicate investigation strategy and status within and outside the team simply does not scale under the high pressure of IT incidents. A new visual paradigm is needed, as visuals are processed by our brains much faster and clearer than reading text.

Investigation Maps™ are a key innovation in smartQED. These maps can be used to visually depict a hierarchy of potential causes for a problem (in the form of a cause tree or Fishbone / Ishikawa diagram – see reference [4]). They additionally provide the capability for users to associate various artifacts such as symptoms (evidence), fault status, notes, actions etc to a specific cause. They are concurrently updatable by users, with automatic merging & notifications, helping to put everyone on the same page quickly.

The benefits of adopting this visual approach are literally transformative! Investigation strategies are clear to all, and the in-context information reduces confusion and greatly enhances shared understanding, leading to much faster resolution of incidents.



As shown in the sample map above, fault status of each cause (i.e., whether cleared or confirmed) is indicated using green and red colors in the Fishbone, so that the overall status can be quickly seen by everyone. Updates to these investigation maps can be pushed to traditional issue tracking tools in time sequence for reporting purposes.

Automated Recommendations from Historical Data

Investigation maps in smartQED enable investigators to easily associate the evidence (e.g., problem symptoms/ observed anomalies) and specify the actions taken for investigation and/or resolution. This historical data is automatically analysed using our proprietary machine learning algorithms to generate recommendations for future problems, based on the problem symptoms.

When a new problem is detected by IT monitoring tools, the symptoms can be sent to smartQED which invokes its Recommendation Engine to identify similar problems that were resolved earlier. The matching problems are further analysed to prescribe likely causes and actions in the form of a Suggested Investigation Map. This map provides a valuable starting point for a new investigation, reducing the need to involve dozens of people from the IT teams.

Crowdsourcing of Problem-Solving Knowledge

Last but not the least, we have designed smartQED to be able to anonymously leverage knowledge from its user community and web forums. The immense popularity of sites like stackoverflow.com and other public forums demonstrates that IT engineers are always looking to share problem-solving knowledge and learn from others.

In smartQED we can analyse and index knowledge from public forums as well as from problems resolved earlier (our suggestions fully respect user data privacy). This enables effective leveraging of ‘crowd wisdom’, and serves to significantly reduce the resolution time (MTTR) of IT incidents and minimizes resources needed to investigate them. Newer members in IT teams are up-leveled and empowered by the expert knowledge from the community, especially for emerging technologies.

4. References

1. Costco's Thanksgiving Day Website Crash Cost It Nearly $11M: https://www.thestreet.com/.amp/technology/costco-thanksgiving-day-website-crash-cost-it-nearly-11million-15185344
2. Internal documents show how Amazon scrambled to fix Prime Day glitches: https://www.cnbc.com/2018/07/19/amazon-internal-documents-what-caused-prime-day-crash-company-scramble.html

3. Cost of Data Center Outages, Ponemon Institute Research Report: https://www.vertiv.com/globalassets/documents/reports/2016-cost-of-data-center-outages-11-11_51190_1.pdf
4. Fishbone / Ishikawa diagram: https://en.wikipedia.org/wiki/Ishikawa_diagram