Production Touch Safety

# Background

Service reliability and quality are at the heart of customer experience and remain key indicators for running mission critical workloads on Azure. Azure has seen several service outages in the past due to erroneous human triggers, DRI fatigue and unsafe production touches. Additionally, operational safety and scaling are vital to run Azure services in sovereign environments and maintain operational efficiency.

# Current Production Touch Landscape

Azure cloud & service build-out and operations go through multiple phases before services are stable and running efficiently. Touches to production occur at every phase of the production pipeline. While a common set of platforms and interfaces exist today for service manageability and reliability, services have also built in-house tools and scripts to manage and operate their service infrastructure. **The common platforms and interfaces do not provide enough controls to ensure safety of the touches causing LSIs and / or CRIs**. In most cases, custom tools do not provide safety validations and guard rails and have caused live site incidents and, in some cases, outages.

The goal of this document is to:

* Define set of principles and practices that a safe platform should provide to make a production touch safe
* Define a safety maturity model and the minimum maturity level for production touches to occur via safe platform
* Define touch classification to allow risk assessment and minimum safety capabilities to be applied by the safe platform
* Identify the platforms and interfaces that should be standardized for every production touch pipeline

# Definitions

|  |  |
| --- | --- |
| **Term** | **Description** |
| Production Touch | Operations that are either user triggered or service-initiated on a production target / resource. In general, these touches can fall into Create, Read, Update or Delete categories. |
| Unsafe Production Touch | Production touches without safety controls / validations that could cause unintended impact to the service like service disruption, outage or customer impacting incident. |
| Critical Resource | Production resources that directly affect the availability, data correctness, performance or functionality issues for the service. There must be a tag in an authoritative system used by the service.  E.g. Storage Account with ACLs for compute access. |
| Non-Critical Resource | Production Resource that does not affect the availability, data correctness and / or performance/ functionality of the service. OR  A non-production Resource. |

# Safe Platform

A Safe Platform must meet the principles below and provide platform capabilities to validate touch safety, each building upon the previous:

* ***Change Classification*** – Provide the ability for system verified change / operation classification (to meet the standard taxonomy i.e. identify operation classes (like read vs. write)) and assign risk level to the change.
* ***Access Controls and Security*** – Enforce granular RBAC and least privilege principles following the [Critical Infrastructure Production Access Standard - SIDN principles](https://microsoft.sharepoint-df.com/:w:/r/teams/ComputeFUNdamentals/OneFleet%20Foundation%20Security/OneFleet%20Foundation%20Standards/Critical%20infrastructure%20production%20access%20standard%20for%20Azure.docx?d=wf2e1e4065b7c487d865c6dacea54a600&csf=1&web=1&e=x5HKnh).
* ***Audited and Attributed Operations*** – Provide auditing and attribution of the touch/change to the identity that requested the activity. Ideally, provide correlation for the attribution to occur through the entire chain of entities involved, i.e. from the original identity to the safe platform to the targeted resource. Push logs into Production Touch Dashboard and FCM.
* ***Managed Risk Review*** – Require review and approval prior to the change. Integration with BoQ / R2D / SafeFly process for risk assessment and approval to ensure the appropriateness, validity & impact of the change / operation. Maintain the integrity of the reviewed and approved changes to production.
* ***Safety Pre-Checks*** – Provide the ability to define and perform pre-checks and validations (depending on the targeted resource) to ensure reduced inadvertent impact.
* ***Safety Policies*** – Protection of resources’ availability by providing default policies for rate and concurrency limits for change execution.
* ***Health Integration*** – Provide the ability for pre- and post- health checks to evaluate the health of the resource before proceeding with the next set of change / operations.
* ***Testing and Safe Deployment Practices*** – Provide capability for testing of the code to validate the operations are doing what is desired. Ensure rollout of change through defined SDP channels such as non-prod, Beta, Canary, regional resilience and prod environments with automated tests, bake time and rollback.
* ***Rollback*** – Ability to revert to the previously known good state if something was not completed successfully.
* ***CCOA Support*** *–* Ability to support no fly periods and meet the requirement of approval process for any change during no-fly

# Platform Capability for Safety

Common platforms may not start safe but, with effort to bridge any feature gaps, can be considered safe over time. Below is the proposed capability model to verify that a platform provides the required safety capabilities to make touches safer:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Capability | Change Classification | Access Control & Security | Audited & Attributed | Managed Risk Review | Safety Prechecks | Safety Policies | Health Integration | Testing & SDP | Rollback | CCOA Support |
| Audited |  |  |  |  |  |  |  |  |  |  |
| Reviewed |  |  |  |  |  |  |  |  |  |  |
| Safe |  |  |  |  |  |  |  |  |  |  |
| Safe at Scale |  |  |  |  |  |  |  |  |  |  |
| Safe w/ Rollback |  |  |  |  |  |  |  |  |  |  |

Safe Standard platform(s) must allow execution of the operation based on the risk evaluation noted in the section below to validate the safety requirements commensurate with the touch risk level – if the safety requirements for the associated risk level is not met the touch operation / change must be blocked.

# Touch Guidance - Classification and Risk Evaluation

Below is the guidance on the minimum set of safety requirements (i.e. Safety Capability) commensurate with the risk level of the operation /touch that the standard platform must enforce on production touches :

| **Touch Type** | **Scope** | **Resource Criticality (Risk??)** | **Description** | **Touch Risk Level** | **Minimum Safety Capability** |
| --- | --- | --- | --- | --- | --- |
| Create | Single Instance / Host | Non Critical | Creation of a single resource, a row of data etc. Write operations that are additive in nature. Build out/capacity addition. | Low | N/A |
| Single Instance or  Multi-region | Critical | Creation operation on a critical resource that may have an impact on the state of the system Read operation for customer content. | Medium | Reviewed |
| Read | Single Instance / Host or Multiple Instances | - | Read operations Operation that allows running diagnostic collection of metrics / telemetry, read state of the resource etc. | Low | N/A |
| Single Instance / Host | - | Read operation to get a secret from a secret vault. Read operation for customer content. | Medium | Reviewed |
| Update | Single Instance / Host | Non Critical | Write operations that are additive in nature | Medium | Reviewed |
| Single instance / Host | Critical | Update of customer content,  Modify state / behavior of a service or a resource.  Metadata changes e.g. billing info | High | Safe at Scale |
| Multiple instances in the same region/AZ or Fault domain | Non Critical | Write operations that are additive in nature | Medium | Reviewed |
| Multiple instances in the same region/AZ or Fault domain | Critical | Update of customer content,  Modify state / behavior of a service or a resource.  Metadata changes e.g. billing info. | High | Safe at Scale |
| Multiple instances across fault domain | Non Critical | Write operations that are additive in nature | High | Safe at Scale |
| Multiple instances across fault domains | Critical | Write operations that are additive in nature or Update operation that changes system/Global/infra configs, Network/state changes | Extreme | Safe w/ Rollback |  |
| Delete | Single Instance / Host or Multiple instances in the same region/AZ or  Multi-Region or across fault domains | Non Critical | Any delete for 1 or more instance of data, config, metadata, infra, decomm, security settings, billing info or Cross AZ/Fault domain changes | High | Safe at Scale |  |
| Single Instance / Host or Multiple instances in the same region/AZ or  Multi-Region or across fault domains | Critical | Any delete for 1 or more instance of data, config, metadata, infra, decomm, security settings, billing info or Cross AZ/Fault domain changes | Extreme | Safe w/ Rollback |  |

|  |  |  |  |
| --- | --- | --- | --- |
| ~~Risk Level~~ | ~~Touch Type~~ | ~~Description~~ | ~~Minimum Safety Requirements~~ |
| ~~Low risk~~ | ~~Read only~~ | ~~Read operations~~  ~~Operation that allows running diagnostic collection of metrics / telemetry, read state of the resource etc.~~ | ~~N/A~~ |
| ~~Medium risk~~ | ~~Read Secrets / Customer Content~~  ~~Create / Insert / Add resources~~  ~~Delete of single non-critical resource~~ | ~~Read operation to get a secret from a secret vault.~~  ~~Read operation for customer content.~~  ~~Write operations that are additive in nature~~  ~~Build out/capacity addition~~ | ~~Managed Risk Review Safety Policies~~ |
| ~~High risk~~ | ~~Update for service / resource state change~~  ~~Update /delete of metadata changes~~  ~~Delete/ Decom of single~~ [~~critical resource~~](#_Critical_resource)  ~~Delete / Decom of multiple non-critical resources~~ | ~~Update of customer content,~~  ~~Modify state / behavior of a service or a resource.~~  ~~Metadata changes e.g. billing info~~ | ~~Managed Risk Review~~  ~~Testing + SDP~~  ~~Safety Prechecks~~  ~~Health Integration~~  ~~Rollback~~ |
| ~~Extreme risk~~ | ~~Update system configs Infrastructure changes~~  ~~Delete / Decom of multiple critical resources~~ | ~~Update operation that changes system/Global/infra configs, Network/state changes~~  ~~Any delete of more than 1 instance of data, config, metadata, infra, decomm, security settings, billing info~~  ~~Cross AZ/Fault domain changes~~ | ~~Managed Risk Review~~  ~~Testing + SDP~~  ~~Safety Prechecks~~  ~~Health Integration~~  ~~Rollback~~ |

# Production Touch Pipelines and Recommended Standards

The production touch pipeline refers to a collection of activities related to the service lifecycle phases. In other words, as service / technology is being built out, deployed, managed, etc., there is often a collection of activities that require production touches. For each of these pipelines, a collection of platform interfaces and tools are being used by the engineers – many of these are considered common due to their broader adoption.

Below is our proposal for recommendations to restrict the platform interfaces / tools for each of the service lifecycle pipelines. The goal is to push the recommended platforms / tools to meet the safety requirements noted in the earlier section to become the Safe Standard.

| Pipeline | Category | Description | Recommended Platform Interface / Tool | |
| --- | --- | --- | --- | --- |
| Below ARM services | Above ARM services |
| Buildout | N/A | Establishing technology to allow service deployment. | [AzCIS](https://eng.ms/docs/products/azcis-platform/getting-started/what-is-cis) | [AzCIS](https://eng.ms/docs/products/azcis-platform/getting-started/what-is-cis) / [EV2](https://ev2docs.azure.net/getting-started/overview.html) |
| Deployment | Code Deployment | Planned changes to the production service’s code, resources / infrastructure. | [AzDeployer](https://eng.ms/docs/cloud-ai-platform/azure-core/one-fleet-platform/onedeploy-absingh/onedeploy-azdeployer/azdeployer-v2/overview/what-is-azdeployer-v2) | [EV2](https://ev2docs.azure.net/getting-started/overview.html) |
| Config Deployment | Planned production touches to update the configuration code for the services’ resources. | TBD  [GA](https://eng.ms/docs/products/geneva/actions/overview) (discussion in progress)  [OaaS](https://eng.ms/docs/cloud-ai-platform/azure-core/azure-networking/sdn-dbansal/sdn-buildout-and-deployments/sdn-fundamentals/tsgs/deploymenttsgs/rnm-deployment/oaas-orchestration-as-a-service/oaas-orchestration-as-a-service) (discussion in progress)  [ConfigGuard](https://eng.ms/docs/cloud-ai-platform/azure-edge-platform-aep/aep-health-standards/sky/microsoft-configguard/configguard/docs/sdk) | [Skylarc](https://eng.ms/docs/cloud-ai-platform/azure-core/one-fleet-platform/one-fleet-efficiency-team/teams/foundation/teams/skylarc) / ECS  [GA](https://eng.ms/docs/products/geneva/actions/overview) (discussion in progress)  [OaaS](https://eng.ms/docs/cloud-ai-platform/azure-core/azure-networking/sdn-dbansal/sdn-buildout-and-deployments/sdn-fundamentals/tsgs/deploymenttsgs/rnm-deployment/oaas-orchestration-as-a-service/oaas-orchestration-as-a-service) (discussion in progress)  [ConfigGuard](https://eng.ms/docs/cloud-ai-platform/azure-edge-platform-aep/aep-health-standards/sky/microsoft-configguard/configguard/docs/sdk) |
| Data Deployment | Deploys data folders by automatically detecting new data versions from source | [ConfigImageBuilder](https://eng.ms/docs/products/autopilot/deployment/troubleshoot/building/configimagebuilder-data-deployment) | - |
| Diagnostics  and Telemetry | N/A | Out-of-band production operations / touches to generate system / service telemetry for system diagnosis in the event of performance issues or failures or incidents. | [Fleet Diagnostics](https://eng.ms/docs/cloud-ai-platform/azure-edge-platform-aep/aep-health-standards/fundamentals/fleet-diagnostics/fleet-diagnostics) | [Fleet Diagnostics](https://eng.ms/docs/cloud-ai-platform/azure-edge-platform-aep/aep-health-standards/fundamentals/fleet-diagnostics/fleet-diagnostics) |
| Mitigation Click / Scripted  Touches | DRI Touches | Out-of-band production touches to remediate or mitigate an incident or an issue with a known fix as a corrective measure. | [Geneva Actions (GA)](https://eng.ms/docs/products/geneva/actions/overview) (operations on a single resource)  [Orchestration as a Service (OaaS)](https://eng.ms/docs/cloud-ai-platform/azure-core/azure-networking/sdn-dbansal/sdn-buildout-and-deployments/sdn-fundamentals/tsgs/deploymenttsgs/rnm-deployment/oaas-orchestration-as-a-service/oaas-orchestration-as-a-service) (orchestrated operations against multiple resources.)  Repair Agent (SoC)  (More to be added..) | [Geneva Actions (GA)](https://eng.ms/docs/products/geneva/actions/overview) (point operations)  [Orchestration as a Service (OaaS)](https://eng.ms/docs/cloud-ai-platform/azure-core/azure-networking/sdn-dbansal/sdn-buildout-and-deployments/sdn-fundamentals/tsgs/deploymenttsgs/rnm-deployment/oaas-orchestration-as-a-service/oaas-orchestration-as-a-service)  SQL CAS  (More to be added..) |
| Support Touches | Production touches performed by Frontline Support engineers to remediate / mitigate a customer issue associated with a support ticket. | [Azure Support Center (ASC)](https://eng.ms/docs/experiences-devices/webxt/bing-places/azure-maps/azure-maps/azure-maps-team-documentation/engineering/asc-azure-support-center) | [Azure Support Center (ASC)](https://eng.ms/docs/experiences-devices/webxt/bing-places/azure-maps/azure-maps/azure-maps-team-documentation/engineering/asc-azure-support-center) |

# Appendix

# Touch Taxonomy

|  |  |  |
| --- | --- | --- |
| **Touch** | **Manual**  Executed on demand by a user identity | **Automated**  Automated execution by service identity |
| **Click Touch** | User executed a “click operation”  e.g. Azure Portal clicks | N/A |
| **Scripted/Code Touch** | User executed a script/code  e.g. ran a PowerShell script | Automation (orchestrator, etc.) invoked the script/code  e.g. OaaS or Geneva Actions execution of an action in response to a trigger or schedule |

# Resource Taxonomy

|  |  |
| --- | --- |
| Environment | Description |
| Development | Infrastructure, subscriptions and/or resources used to develop / code service and related features / changes. Considered as Non Production. |
| Testing and Staging | Infrastructure, subscriptions and/or resources used to perform tests (manual or automated) to validate new or updated code. Includes Staging and end user /client test environments. Considered as Non Production. |
| Production | Infrastructure, subscriptions and / or resources used to host live service where end customers would directly interact with the service.  Resources that host customer content. |
| Test-In-Production | Infrastructure hosts that are rotated to be in Production and out (in transient state). Considered as Production. |

# Critical Infrastructure Production Access Standard - SIDN Principles

Primary objective of the [Critical Infrastructure Production Access Standard](https://microsoft.sharepoint-df.com/:w:/r/teams/ComputeFUNdamentals/OneFleet%20Foundation%20Security/OneFleet%20Foundation%20Standards/Critical%20infrastructure%20production%20access%20standard%20for%20Azure.docx?d=wf2e1e4065b7c487d865c6dacea54a600&csf=1&web=1&e=x5HKnh) is to ensure that Azure services are guarded against threats involving authorized access to production systems and services and eliminate vectors that can be used to compromise the platform. Below are the key SIDN principles from this standard:

* Production access is allowed only for production identities (\*ME and Torus)
* Production access is allowed only through approved devices (like SAW) and approved networks
* Production access is gated with role based, fine grained and time bound access controls

# Manual Touch (Click Touch + Scripted Touch) Scenarios and Recommended Platform / Tools

| Pipeline | Resource | Layering | Specific Scenarios | Recommended Platform |
| --- | --- | --- | --- | --- |
| Diagnostics and Telemetry | Virtual machine | Above ARM | I need to collect process dumps for Monitoring agent | Fleet Diagnostics and Watson |
| Host machine | Below ARM | I need a network packet capture | Fleet Diagnostics and Watson |
| VM and Host Machine | Above and Below ARM | I need to collect a memory dump | Fleet Diagnostics and Watson |
| Virtual machine | Below ARM | I need to collect Guest Agent VM logs | Fleet Diagnostics and Watson |
| Monitoring Agent | Above and Below ARM | I need to check Mon Agent Health | Jarvis, AzSecPack |
| Virtual machine | Above AM | I need to get the VM screenshot | Fleet Diagnostics and Watson |
| Virtual machine | Above ARM | I need to run virtual switch diagnostics for investigating issues that happened during the stop destroy scenario | Fleet Diagnostics and Watson |
| Service Bus | Above ARM | I want to get the scale unit info from ARM ID | Geneva Actions |
| Storage Cluster | Below ARM | I want to get Cluster capacity for a regional account | Geneva Actions |
| Mitigation | Host Machine, | Below ARM | I need to execute a PS script to restart hundreds of host nodes across different fabric clusters. | OaaS |
| Network Resource Provider | Below ARM | I Need to Invoke DeleteDynamicSetting for my feature in all of Prod | OaaS |
| Node | Below ARM | I need to Power on/Power Off/ Power Cycle/ Reboot a Node/Server | Geneva Actions |
| Virtual machine | Below ARM | I need to set up automated fault handling when a customer VM is down due to disk failures | Anvil |
| Host Node | Below ARM | I need to perform automatic VM migration on a node that has lost connectivity to the network. | Anvil |
| Storage Role instance | Below ARM | I need to restart a Storage Role | Geneva Action |
| VMSS | Above ARM | I need to delete VMSS from a resource group in a Prod subscription | Geneva Action (Shared operation) |
| Traffic Manager Profile | Above ARM | I need to delete an Azure Traffic Manager Profile | Geneva Actions (Shared Operation) |
| Storage Account | Above ARM | I need to delete a Storage Account in my subscription, RG | Geneva Actions (Shared Operations) |
| Web Apps | Above ARM | I want to delete a website in my subscription | Geneva Actions (Shared Operations) |
| Service Fabric | Above ARM | I need to reimage PaaS Node | Geneva Actions |
| Managing Portal RG’s | Above ARM | I need to manage ACS clusters and node pools under a subscription. | EV2 + GA |
| OVL SOC | Below ARM | I need to power Cycle SOC node. | Repair Agent. |
| OVL SOC | Below ARM | I need to execute a binary script on a SOC node. | Repair Agent |

# Geneva Actions – Platform Safety

[Harden Geneva Actions Platform Safety.docx](https://microsoft.sharepoint.com/:w:/t/Acisandrunners/ER9JGSP0cu5FsUj4jiQ2NBIBwkXMzw9hXLlMqp5wTGHkzA?e=j4n6gO)

# Capability Matrix for known standard platforms

| **Tool Name** | **Change Classification** | **Access Control + Security** | **Audited + Attributed** | **Managed Risk Review** | **Safety Prechecks** | **Safety Policies** | **Health Integration** | **Testing + SDP** | **Rollback** | **CCOA Support** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **OaaS** | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | No | Yes |
| **Fleet Diagnostics** | No | Yes | Yes | No | No | Yes | Yes | Yes | Yes |  |
| **Ev2** | No | Yes | Yes | Yes | \* | Yes | Yes | Yes | Yes | Yes |
| **Geneva Actions** | No | Yes | Yes | No | \* | Yes | Yes | Yes | No | No |
| **AzDeployer** | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **Skylarc** | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **SQL CAS** | No | Yes | Yes | Yes | Yes | Yes | Yes | No | No | Yes |
| **XDS** | No | Yes | Yes | Yes | \* | Yes | Yes | No | Yes | No |
| **DCM Explorer** | No | Yes | Yes | No | No | No | No | No | No | No |
| **Anvil** | No | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes | Yes |
| **NetAssist** | No | Yes | Yes | Yes | \* | \* | Yes | Yes | Yes | Yes |

\* exists for some instances