Bridging the Safety Gap from Scripts to Full Auto-Remediation

David Mah
Dropbox

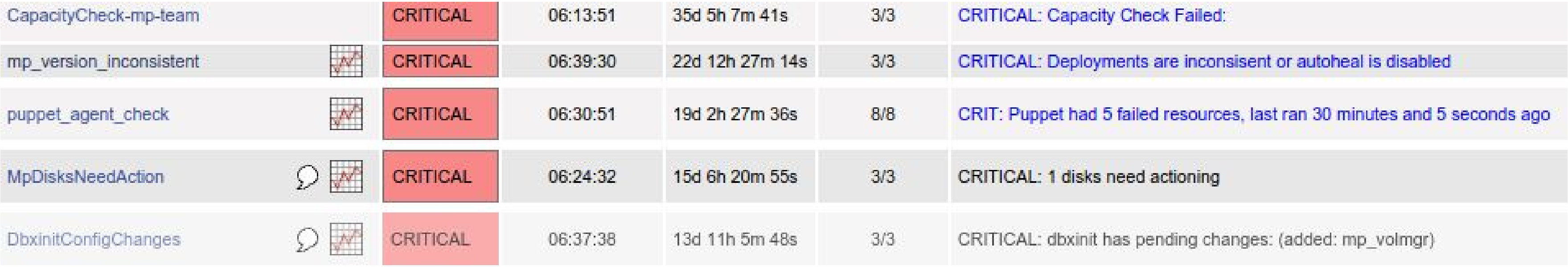
Scary Tasks

Life of Tooling

Naoru Workflow

Dropbox Storage-team
On-call Maintanenence

Dropbox Storage-team On-call Maintanence



CapacityCheck-mp-team		CRITICAL	06:13:51	35d 5h 7m 41s	3/3	CRITICAL: Capacity Check Failed:
mp_version_inconsistent	📈	CRITICAL	06:39:30	22d 12h 27m 14s	3/3	CRITICAL: Deployments are inconsisent or autoheal is disabled
puppet_agent_check	📈	CRITICAL	06:30:51	19d 2h 27m 36s	8/8	CRIT: Puppet had 5 failed resources, last ran 30 minutes and 5 seconds ago
MpDisksNeedAction	💬 📈	CRITICAL	06:24:32	15d 6h 20m 55s	3/3	CRITICAL: 1 disks need actioning
DbxinitConfigChanges	💬 📈	CRITICAL	06:37:38	13d 11h 5m 48s	3/3	CRITICAL: dbxinit has pending changes: (added: mp_volmgr)

...

425 of 425 Matching Service Entries Displayed

Scary Tasks

Scary Tasks

Rebooting Servers

Scary Tasks

Rebooting Servers

Hardware Replacements

Scary Tasks

Rebooting Servers

Hardware Replacements

Reformatting Hard Drives

We need to safely automate these problems away

Scary Tasks

Life of Tooling

Naoru Workflow

Life of Tooling

Life of Tooling

Manual Debug + Manual Fix

Life of Tooling

Manual Debug + Manual Fix

Alert + Manual Fix

Life of Tooling

Manual Debug + Manual Fix

Alert + Manual Fix

Alert + Script to Fix

Life of Tooling

Manual Debug + Manual Fix

Alert + Manual Fix

Alert + Script to Fix

Alert + Auto Run Script

Life of Tooling

Manual Debug + Manual Fix

Alert + Manual Fix

Alert + Script to Fix

Alert + Human Auth Script

Alert + Auto Run Script

Human Authorized Execution

```
A kernel upgrade is pending.

We must reboot the host to get the upgrade

       reboot_host ash-ra9-6a

May this run? [y/n]
```

Why should I NOT run this?

Why should I NOT run this?

We can automate that query

Life of Tooling

Manual Debug + Manual Fix

Alert + Manual Fix

Alert + Script to Fix

Alert + Human Auth Script

Alert + Auto Run Script

Life of Tooling

Manual Debug + Manual Fix

Alert + Manual Fix

Alert + Script to Fix

Alert + Human Auth Script

Alert + Auto Run Script

Automate verification of safety

Why should I NOT run this?

Scary Tasks

Life of Tooling

Naoru Workflow

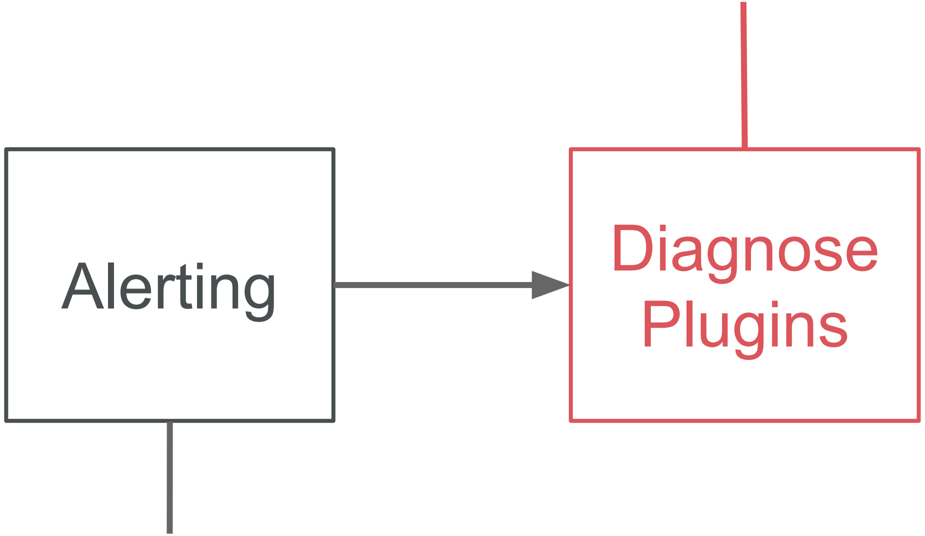
Alerting

Alerting:

- Automate away ‘finding problems’

DiagnosePlugins:

- Find **root cause**
- Decide action to take



Alerting

Diagnose Plugins

Alerting:

- Automate away ‘finding problems’

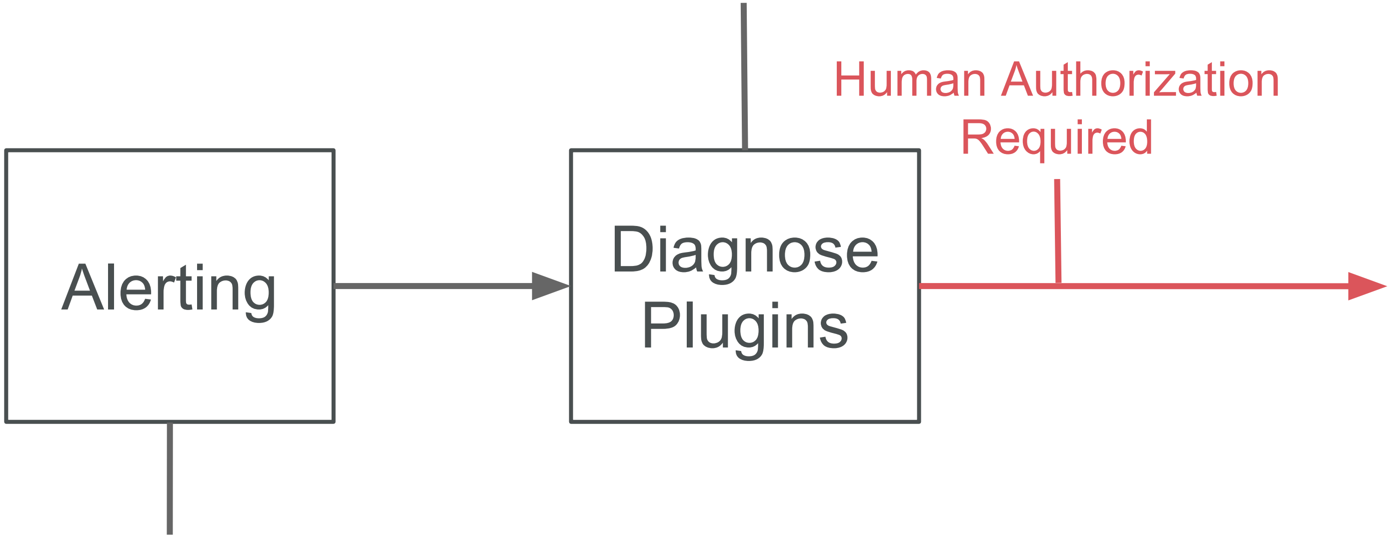
DiagnosePlugin Example

SSH Unreachable Alert?

1. If SSH actually works → do nothing
2. If repeated reboot history → deallocate
3. Check out-of-band IPMI console
    1. If no response → reboot
    2. If (initramfs) shell → deallocate
    3. If cpu soft lockup → reboot

DiagnosePlugins:

- Find root cause
- Decide action to take



Alerting

Diagnose Plugins

Human Authorization Required

Alerting:

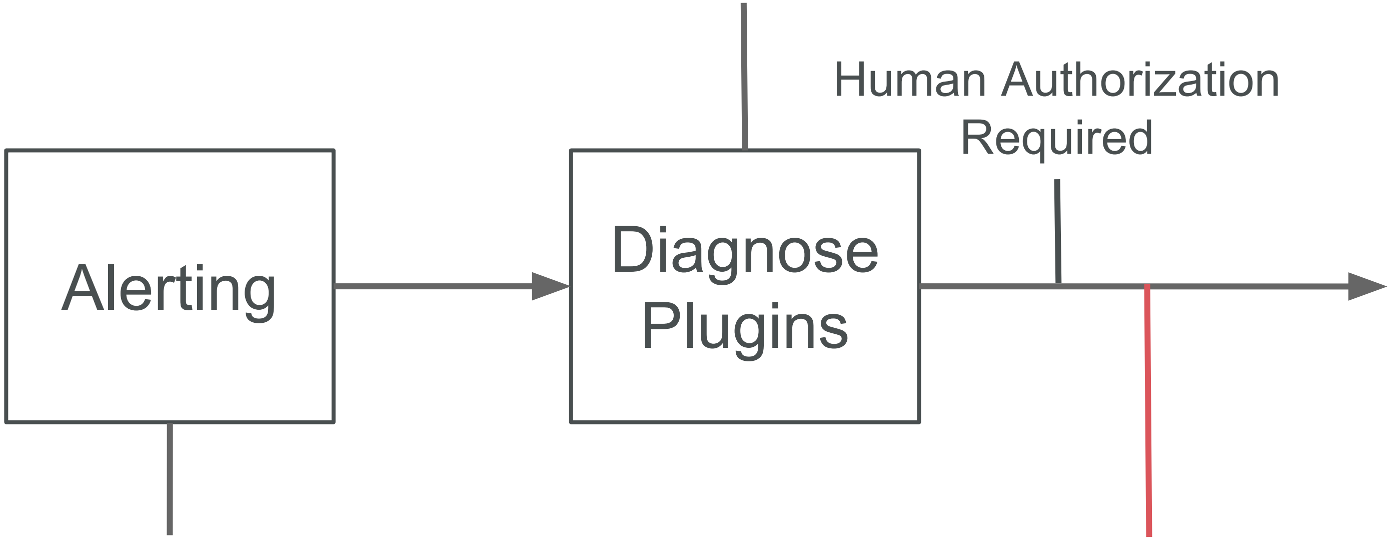
- Automate away ‘finding problems’

Human Authorized Execution

```
A kernel upgrade is pending.

This remediation will reboot the host (zookeeper participant)

       reboot_host ash-ra9-6a

This is safe because
*  We have a rate limit of 1 participant per 5 minutes
*  We check for 4/5 participant availability

May this run? [y/n]
```

DiagnosePlugins:

- Find root cause
- Decide action to take



Human Authorization Required

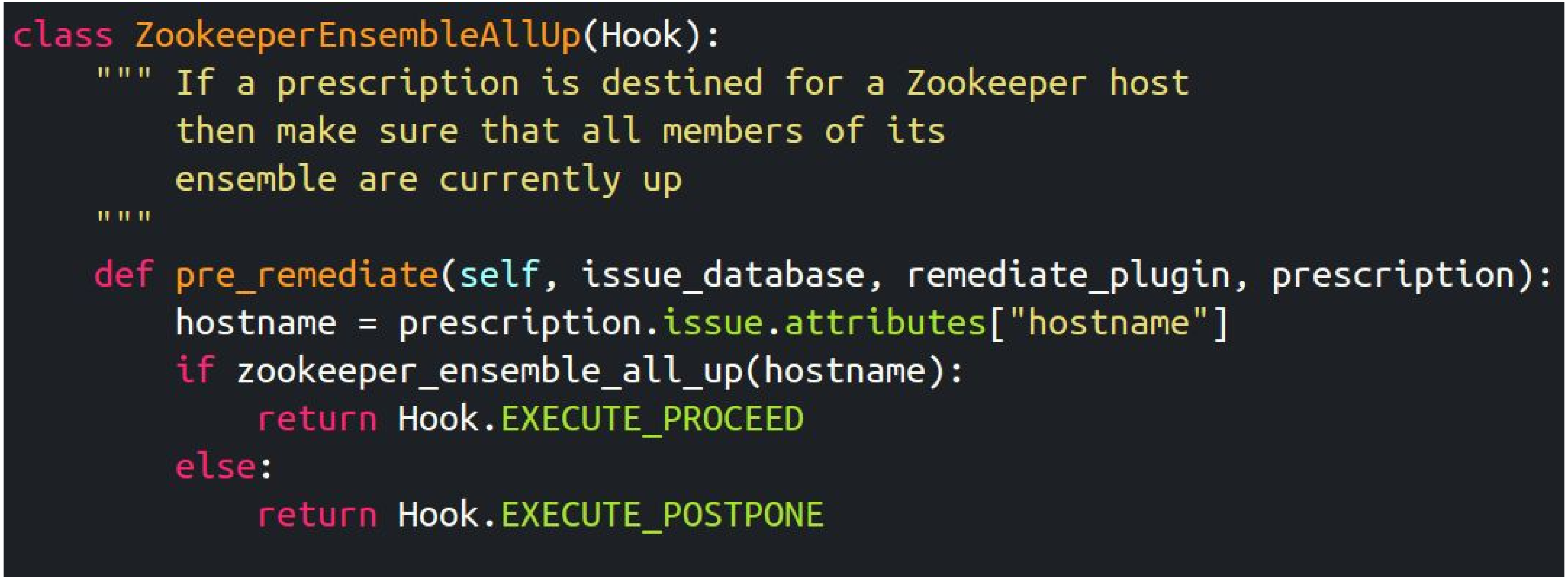
Alerting:

- Automate away ‘finding problems’

pre_remediate hooks:

- Possibly POSTPONE the remediate plugin
  - Rate limits
  - Sanity checks

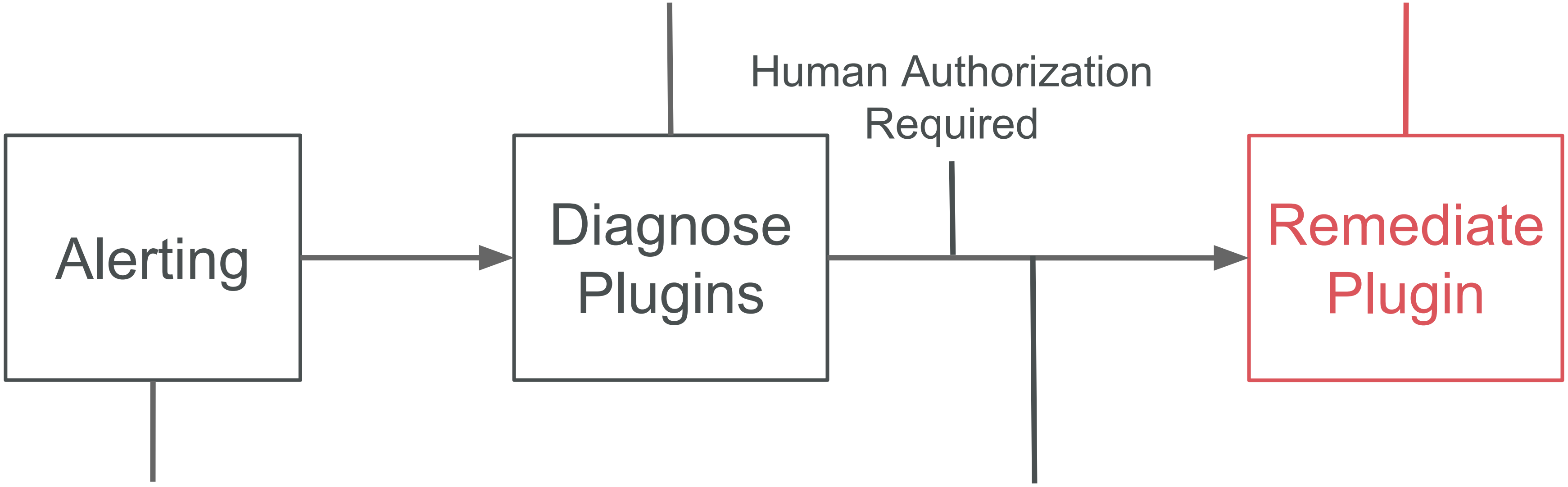
Example of a Sanity Check



```python
class ZookeeperEnsembleAllUp(Hook):
    """ If a prescription is destined for a Zookeeper host
        then make sure that all members of its
        ensemble are currently up
    """
    def pre_remediate(self, issue_database, remediate_plugin, prescription):
        hostname = prescription.issue.attributes["hostname"]
        if zookeeper_ensemble_all_up(hostname):
            return Hook.EXECUTE_PROCEED
        else:
            return Hook.EXECUTE_POSTPONE
```

Example of a Rate Limit

```
Reboot:
    matches:
        remediate_plugin:
            - RebootSSH
            - RebootIPMI
    group_by:
        - service
    limit: 1
    Time_window: 300
    comment: Reboots are 1 per 5 minutes per-service
```



DiagnosePlugins:

- Find root cause
- Decide action to take

RemediatePlugin:

- Actually fix the problem

Human Authorization Required

Alerting:

- Automate away ‘finding problems’

pre_remediate hooks:

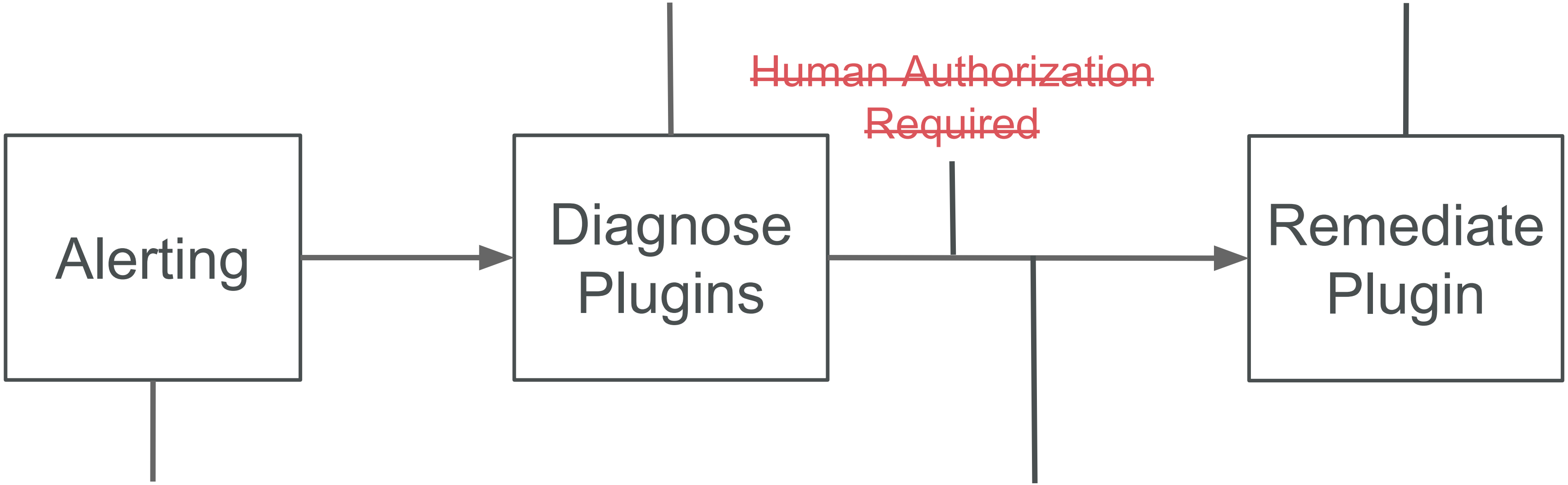
- Possibly POSTPONE the remediate plugin
  - Rate limits
  - Sanity checks

DiagnosePlugins:

- Find root cause
- Decide action to take

RemediatePlugin:

- Actually fix the problem



Alerting

Diagnose Plugins

~~Human Authorization Required~~

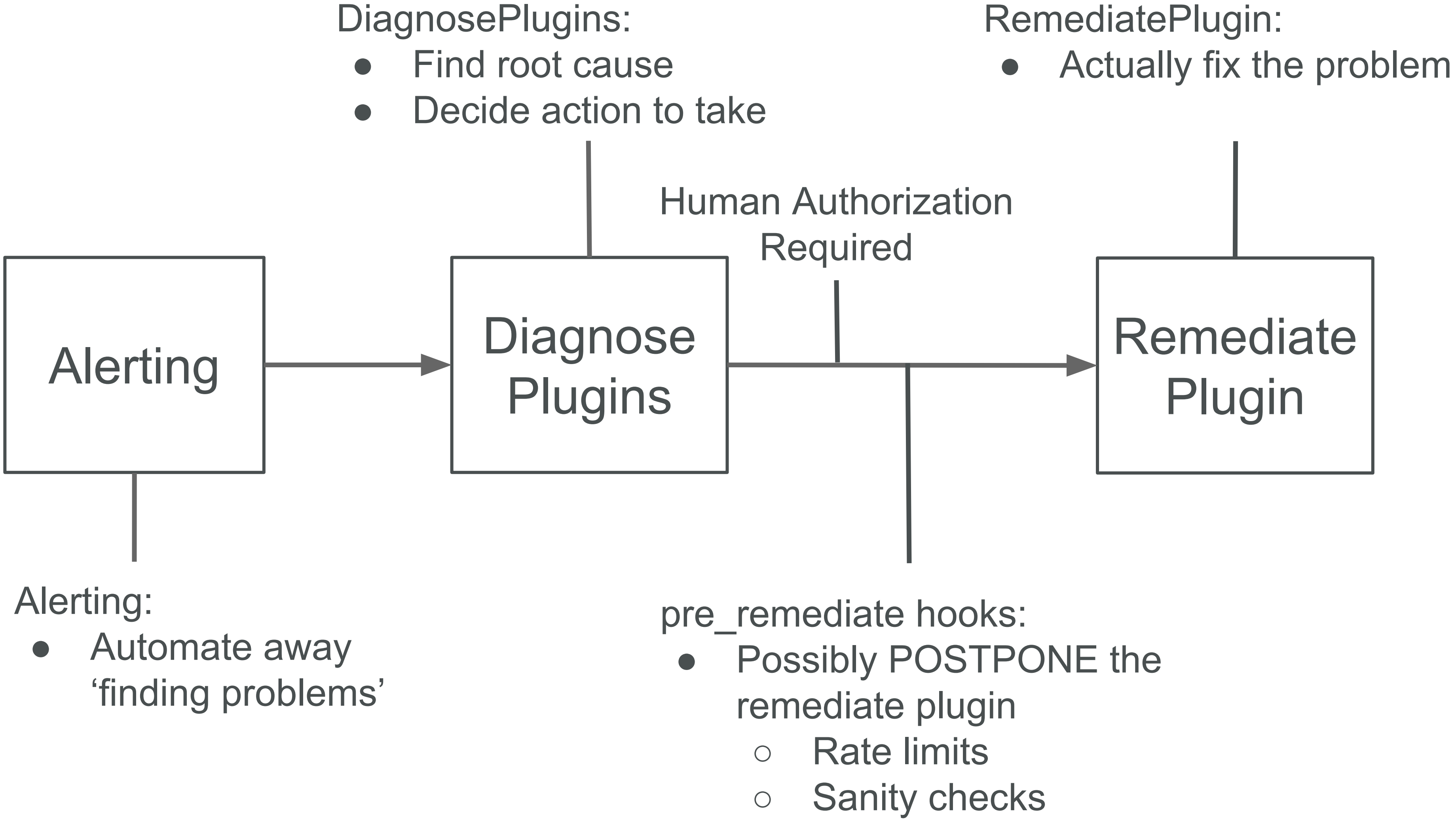
Remediate Plugin

Alerting:

- Automate away ‘finding problems’

pre_remediate hooks:

- Possibly POSTPONE the remediate plugin
  - Rate limits
  - Sanity checks



Alerting

Diagnose Plugins

Human Authorization Required

Remediate Plugin

Alerting:

- Automate away ‘finding problems’

DiagnosePlugins:

- Find root cause
- Decide action to take

pre_remediate hooks:

- Possibly POSTPONE the remediate plugin
  - Rate limits
  - Sanity checks

RemediatePlugin:

- Actually fix the problem

Thanks for stopping by!

David Mah
mah@dropbox.com