从疲于奔命“救火”到自动化运维

—— eBay Hadoop集群自动化运维实践

Jing Ge | 葛京
Polo Li | 李健
eBay Cloud Services

About us

Jing Ge

1st

Manager at eBay

Hamburg Area, Germany | Information Technology and Services

Current eBay

Previous Kuehne + Nagel, MSC (Bertling EDI Service & IT GmbH), Tsinghua University Publish

Education Technische Universität Hamburg-Harburg

Send a message

102 connections

https://de.linkedin.com/in/gejing

Contact Info

About us

Polo Li

Hadoop Services CCOE Team Leader - eBay China Development Center

Pudongxin District, Shanghai, China | Computer Software

Current	eBay China Development Center
Previous	eBay China Operation Center
Education	University of Science and Technology of China

View profile as

242
connections

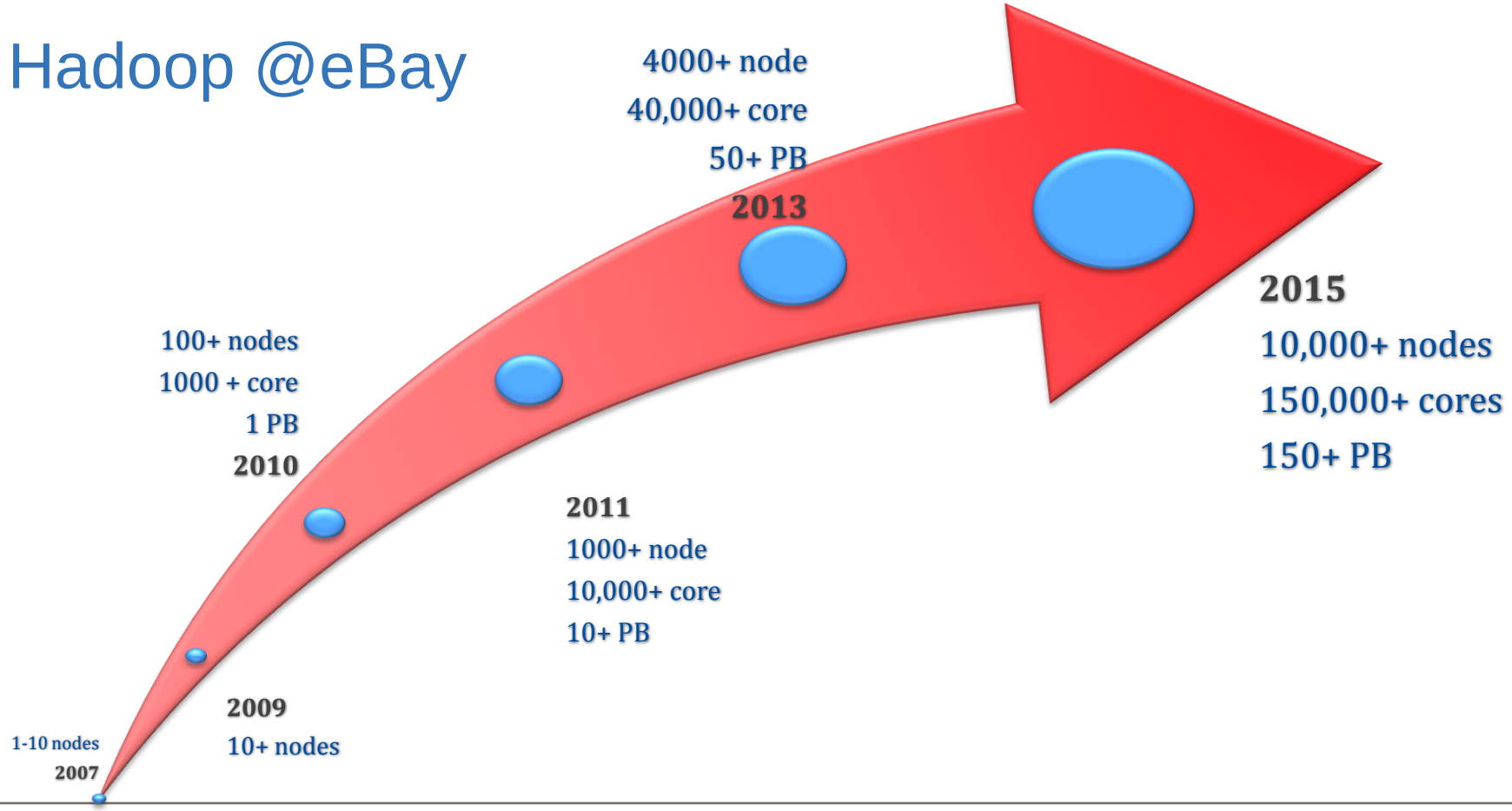
https://cn.linkedin.com/in/polo-li-84477186

Contact Info

Agenda

- **Hadoop @eBay**
- Problem Statement
- What is Hadoop Robot
- Q&A

Hadoop @eBay


4000+ node
40,000+ core
50+ PB
2013
2015
10,000+ nodes
150,000+ cores
150+ PB
100+ nodes
1000 + core
1 PB
2010
2011
1000+ node
10,000+ core
10+ PB
2009
10+ nodes
1-10 nodes
2007

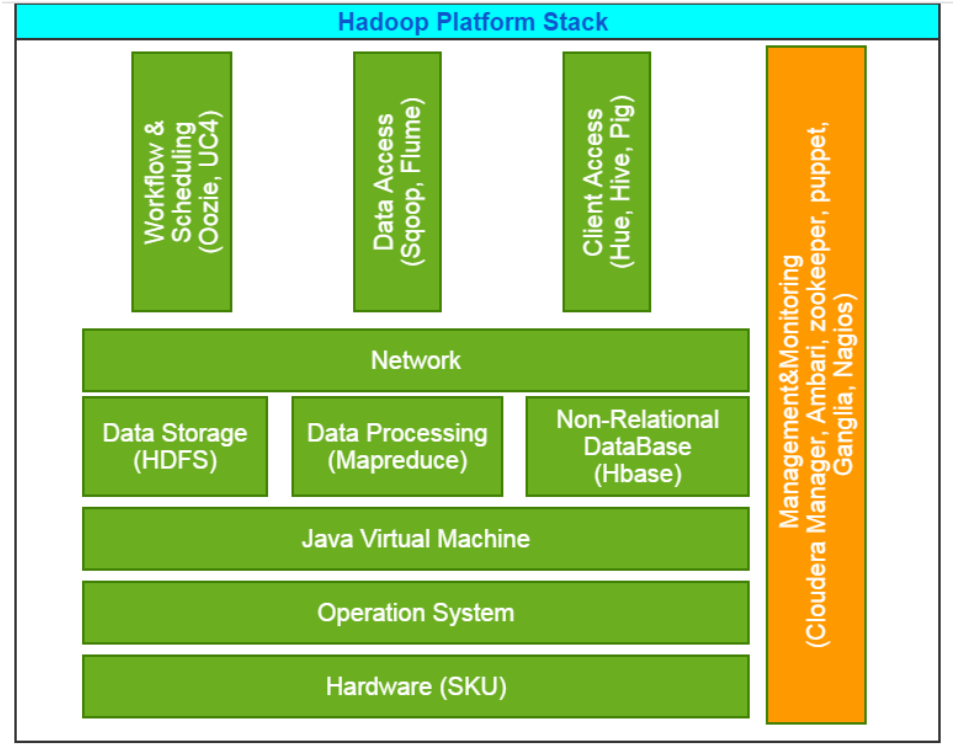
Hadoop @eBay

- 10+ large Hadoop Clusters
- 10,000+ nodes
- 50,000+ jobs per day
- 50,000,000+ tasks per day

Shared vs Dedicated Clusters

- **Shared clusters**
  - Used primarily for analytics of user behavior and inventory
  - Mix of batch and ad-hoc jobs
  - Mix of MR, YARN, Hive, Pig, Cascading, etc.
  - Hadoop and HBase security enabled
- **Dedicated clusters**
  - Very specific use case like index building
  - Tight SLAs for jobs (in order of minutes)
  - Immediate revenue impact
  - Usually smaller than our shared clusters, but still large (600+ nodes)

Hadoop Platform Stack



Team Responsibility

- Full hadoop stack support for Cassini Hadoop (CDH)- hardware up to (and including) the Hadoop/HBase platform itself

- Full hadoop stack support for Analytics Hadoop (HDP) - hardware up to (and including) the Hadoop/HBase platform itself

Daily Work Overview

- Hadoop Maintenance to fix the bad nodes (having disk, nic, cpu, mem, fan, power supply, bmc or the other hardware problems) and keep the live nodes percent >= 98% for all production hadoop clusters
- Keep up with dozens of requests for software and configuration updates/upgrades on the Hadoop/Hbase platform
- Quickly diagnose production problems and rapid response to any Hadoop/Hbase, hardware, os issues.
- Build the new clusters or expand the current clusters
- Monitor all the production hadoop clusters with OS and Hadoop Metrics
- Monitor running jobs performance
- Hadoop Clusters management – HDFS Quota, Queue, Permission, Trash setting, enable audit logs and make any necessary tuning changes to clusters
- Deal with the linux kernel issues
- Deal with JVM issues
- Deal with the oozie && cm mysql db issues
- Linux OS, firmware and hardware upgrades
- Hadoop automation
- 24 * 7 on call support for production hadoop clusters

Agenda

- Hadoop @eBay
- **Problem Statement**
- What is Hadoop Robot
- Q&A

Problem Statements

- Long trouble shooting time
- Bad cluster performance
- Too many different skus, operating systems and metadatas
- Human resource Cost
- Cluster Availability

Traditional Trouble Shooting Pipeline

Step 1
- Check failed application task logs to find out the suspicious hadoop nodes

Step 2
- Check the suspicious hadoop node hardware && system status

Step 3
- Check hadoop metrics and hadoop daemon logs

Step 4
- Check hadoop source code

Victim or Perpetrator ?

Sometimes you think you've found the perpetrators, however it may turn out to be the victim.

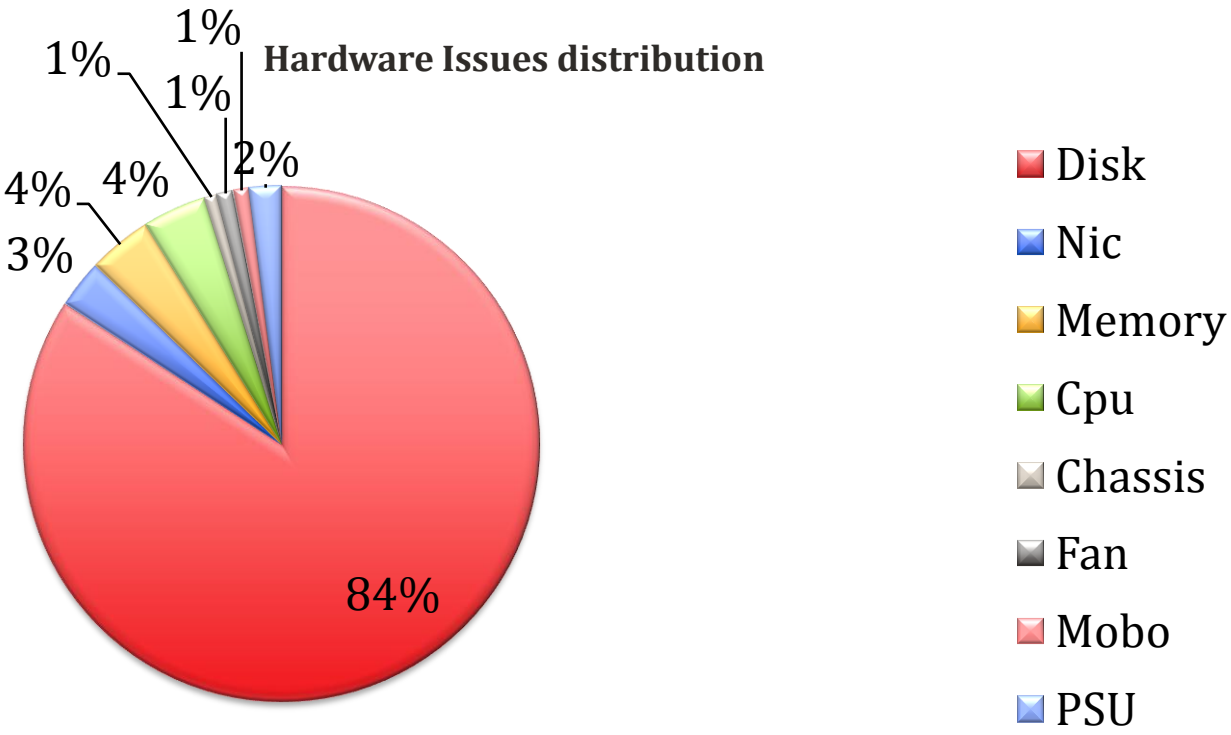
What may impact cluster performance?

- Hardware
- System
- Hadoop
- JVM

Advantages vs Disadvantages

Hadoop can run on cheap(er) hardware but ensuring good performance is a challenge due to less fault-tolerant hardware.

Hardware issues



System issues

- High load
- Node reboot
- Disk full
- Network saturated fully
- OOM
- Kernel bug
- Orphan processes
- …

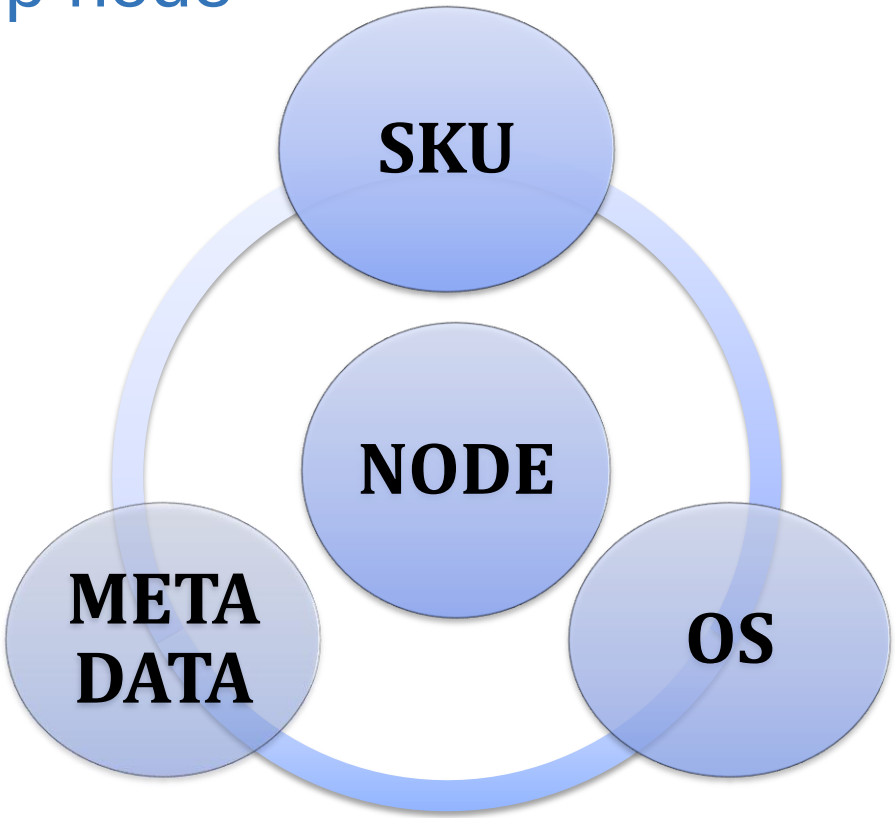
Hadoop Issues

- Hot spot
- Low hdfs locality
- High RPC call queue length
- Hadoop configuration inconsistent issues
- Big resource consuming applications
- Big log/hdfs output applications
- Bad Application scheduling
- …

Premise

- Unhealthy hardware exhibit differences when compared normal, healthy nodes.

About hadoop node



Different SKU

- Cpu core
- Memory size
- Disk number
- Disk size
- Nic Speed
- …

Different OS

- Image profile
- Image version
- Patches
- …

Different Metadata

- **Hadoop Metadata**
  - Daemons
  - Packages
  - Configurations
- **OS Metadata**
  - Services
  - Scripts/Tools
  - Configurations

Human Resource Cost

- Detect node (varies based on admin experiences)
- Node decommission (1-2 Hours)
- Vendor offline remediation (3-5 Days)
- Node reimage（45 Mins）
- Hadoop Installation and OS configuration (15 Mins)
- Node restart (5 Mins)
- Disk burning test (6-8 Hours)
- Node health verification (15 Mins)
- Node recommission (15 Mins)

Cluster Availability

- Improve Availability
  - Increase Live data node percentage
  - Reduce job failures due to bad node

How to make cluster management easier ?

We need to locate unhealthy nodes and offline them as quickly as possible without any manual trouble shooting.

Best practice is that make sure the nodes of the same sku having same

-- Operation System

-- Hadoop and OS Metadata

Hardware Maintenance work is laborious and expensive – automation is a necessity.

Agenda

- Hadoop @eBay
- Problem Statement
- **What is Hadoop Robot**
- Q&A

What is Hadoop Robot

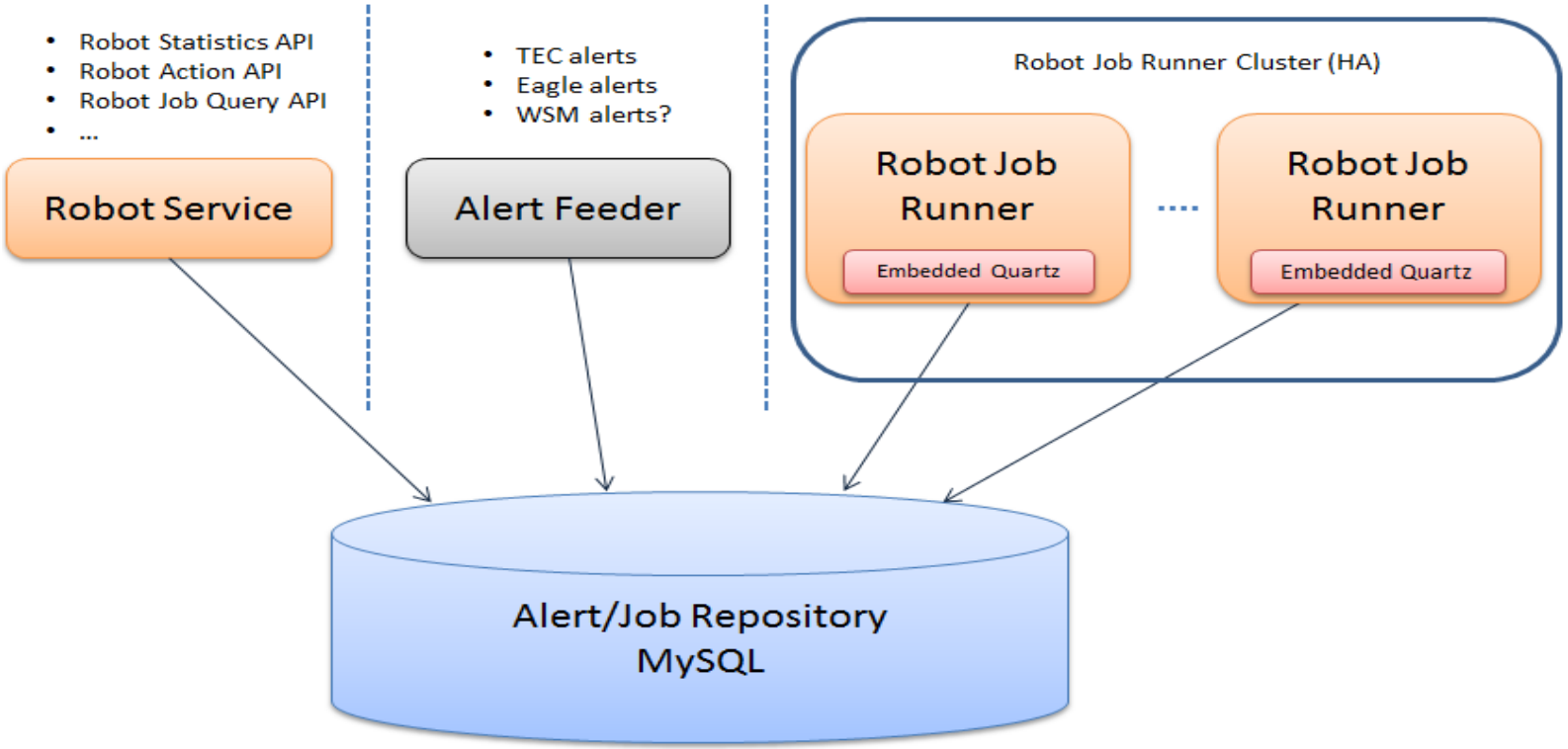
Hadoop Robot is action and remediation center for eBay hadoop clusters:

- End-to-end automated remediation center
- API center for hadoop action and remediation
- Unified Hadoop Admin Console
- Real time maintenance view of Hadoop clusters
- Analytical insights into hardware maintenance data

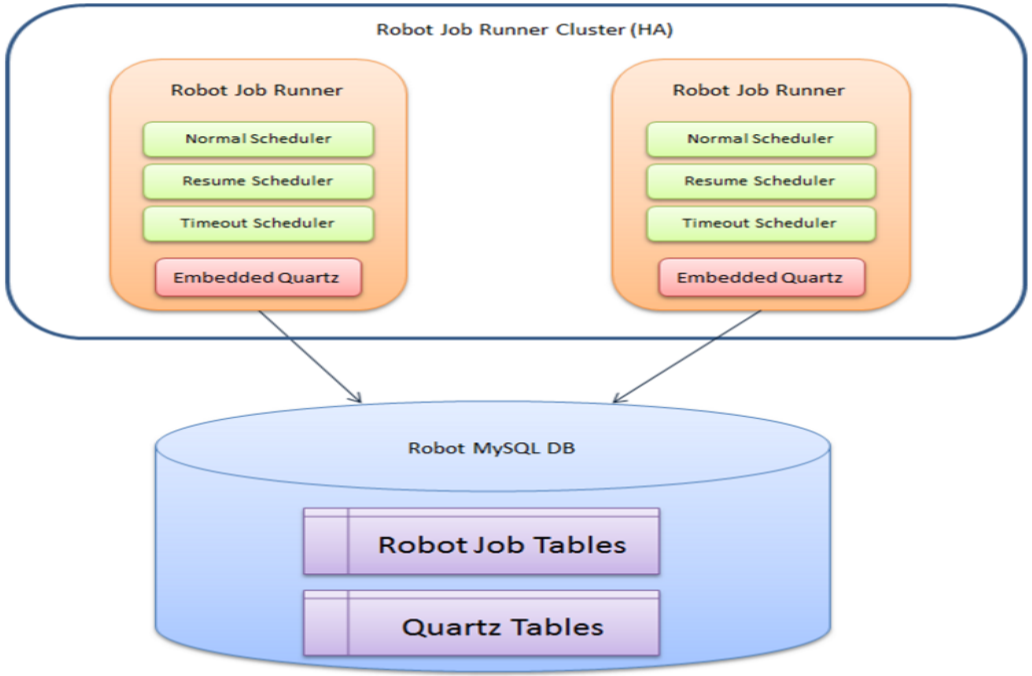
End-to-end Automatic remediation center

- Hardware Maintenance
  - Alert Detection
  - Node Decommission
  - Remediation
  - Node Recommission
- Remove Failed Disk Volume
- Bad Disk Hot Swap
- Hadoop Daemon Restart
- Hadoop Abnormal Job Termination
- Hadoop Cluster Expansion
- …

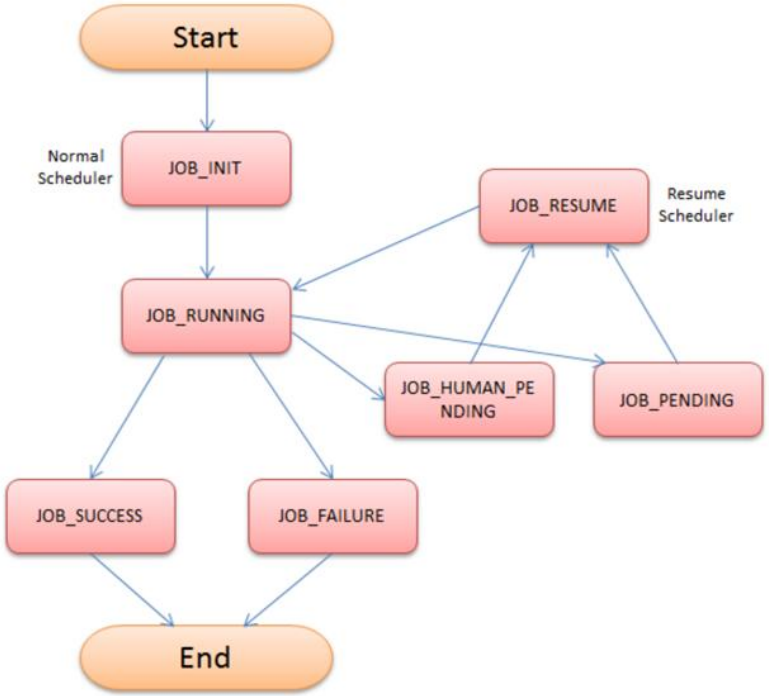
Robot Architecture



Robot Job Runner



Job Status State Machine

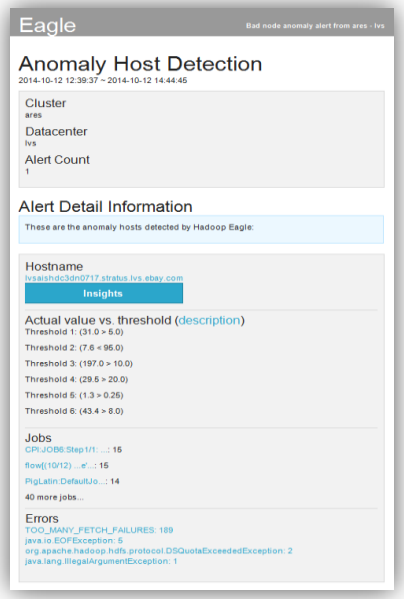


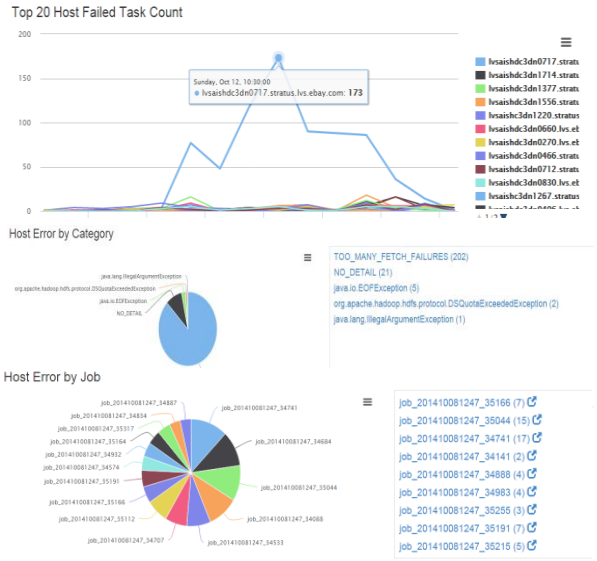
Eagle Alerts

Alerting: Anomaly Detection & Alerting

Insight: Task failure drill-down

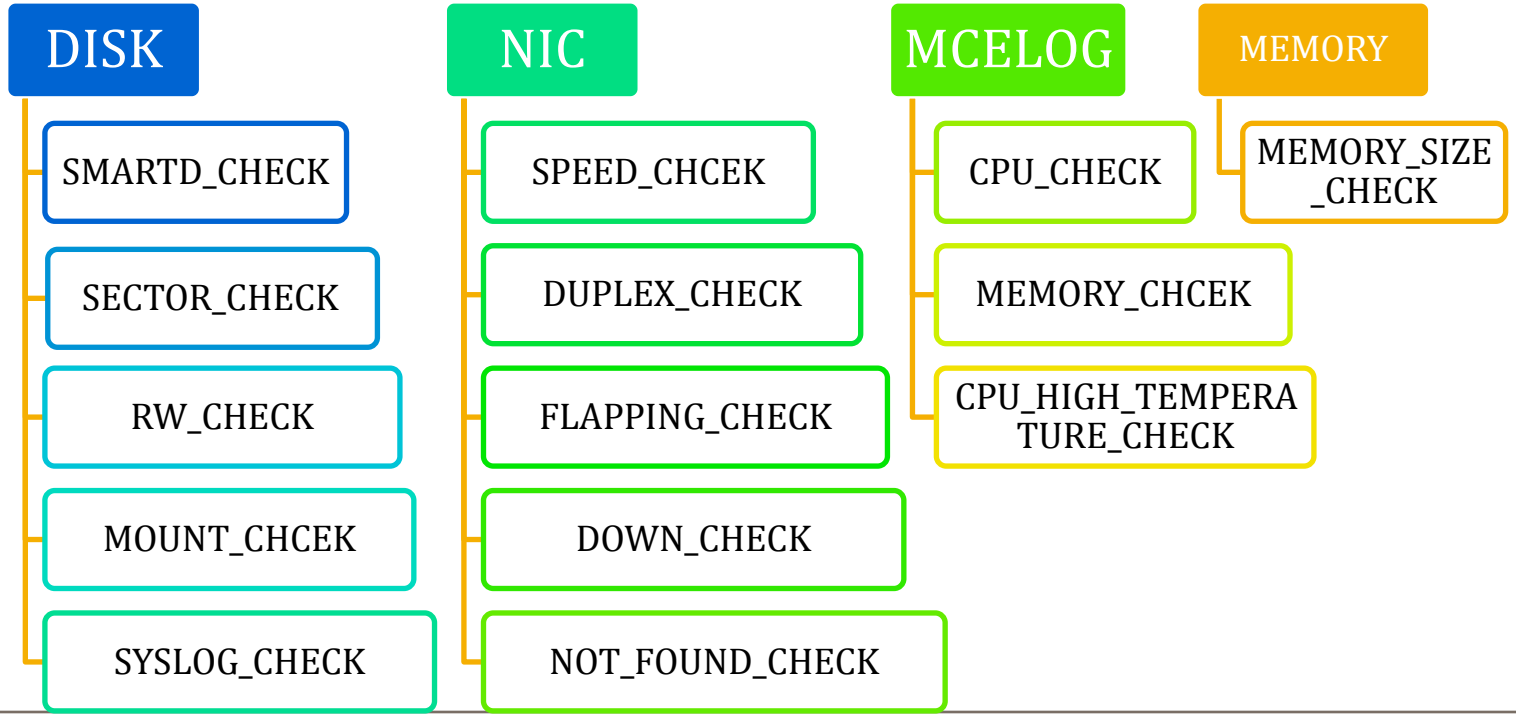
Insight: Task failure drill-down





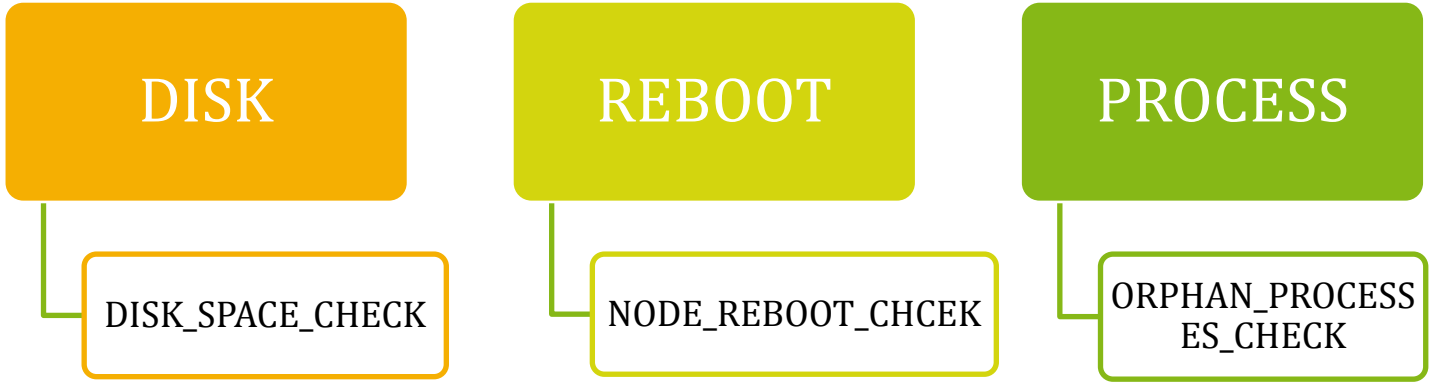
Hardware and System Alerts

- **Hardware**

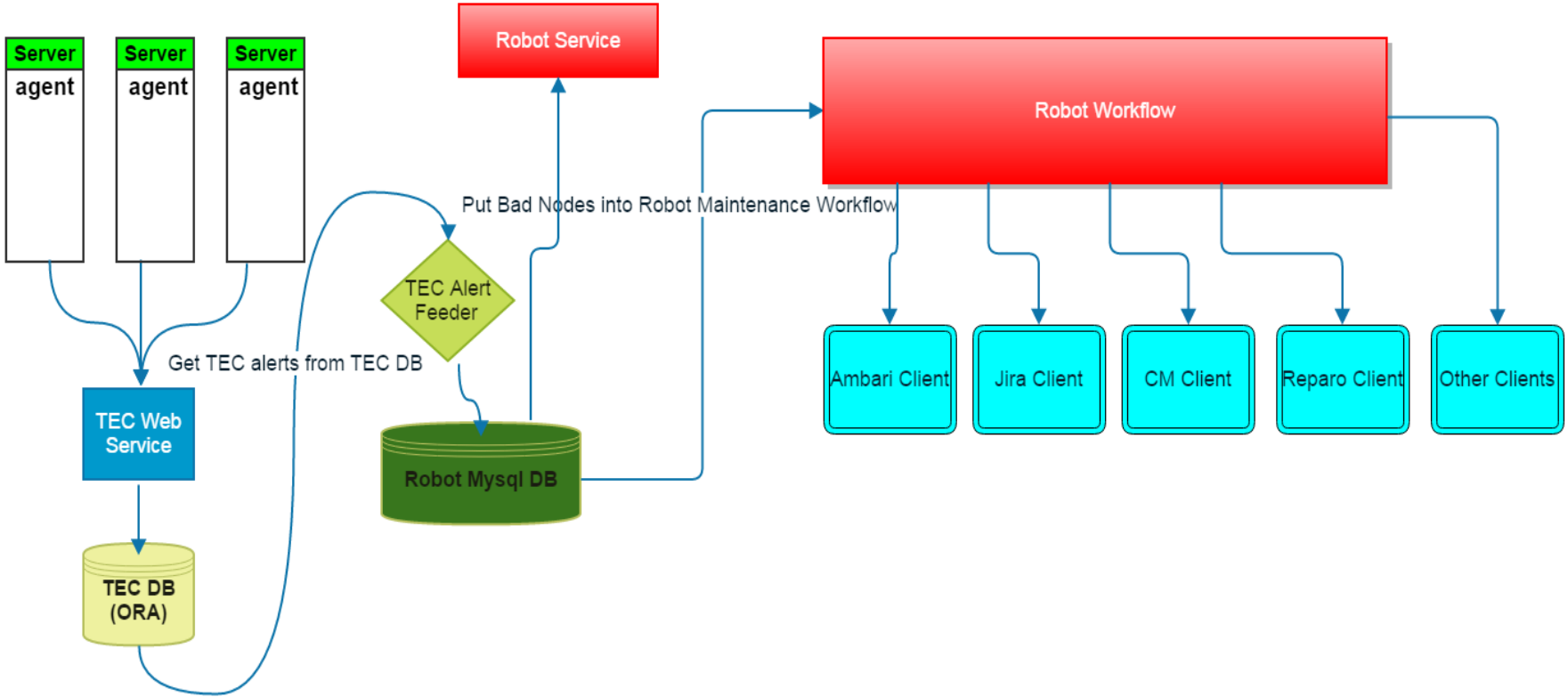


Hardware and System Alerts

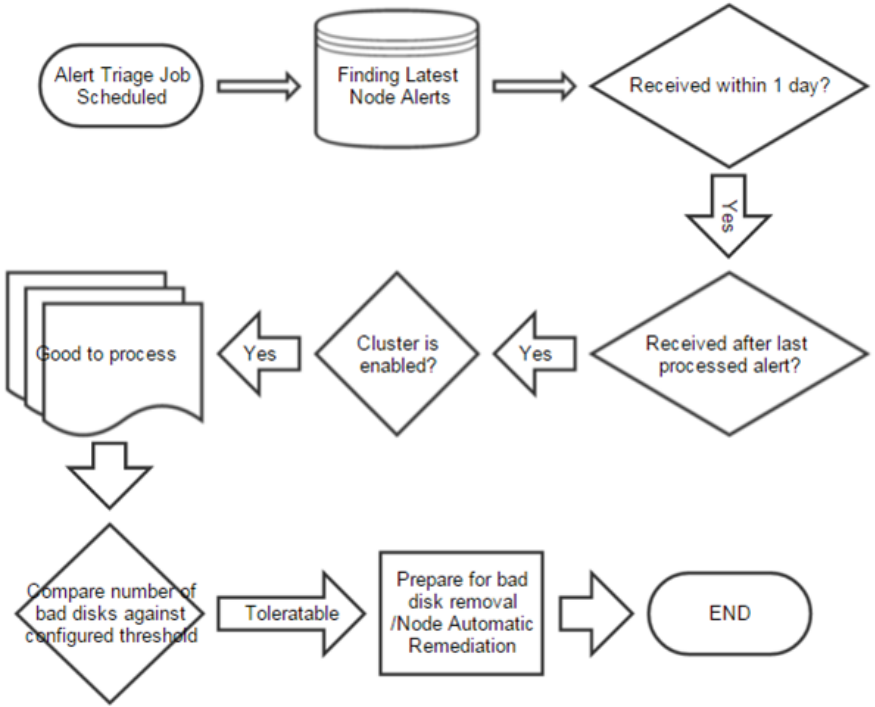
- System



Data Flow


Server
agent
Server
agent
Server
agent
Robot Service
Robot Workflow
Put Bad Nodes into Robot Maintenance Workflow
TEC Alert Feeder
Get TEC alerts from TEC DB
TEC Web Service
Robot Mysql DB
TEC DB (ORA)
Ambari Client
Jira Client
CM Client
Reparo Client
Other Clients

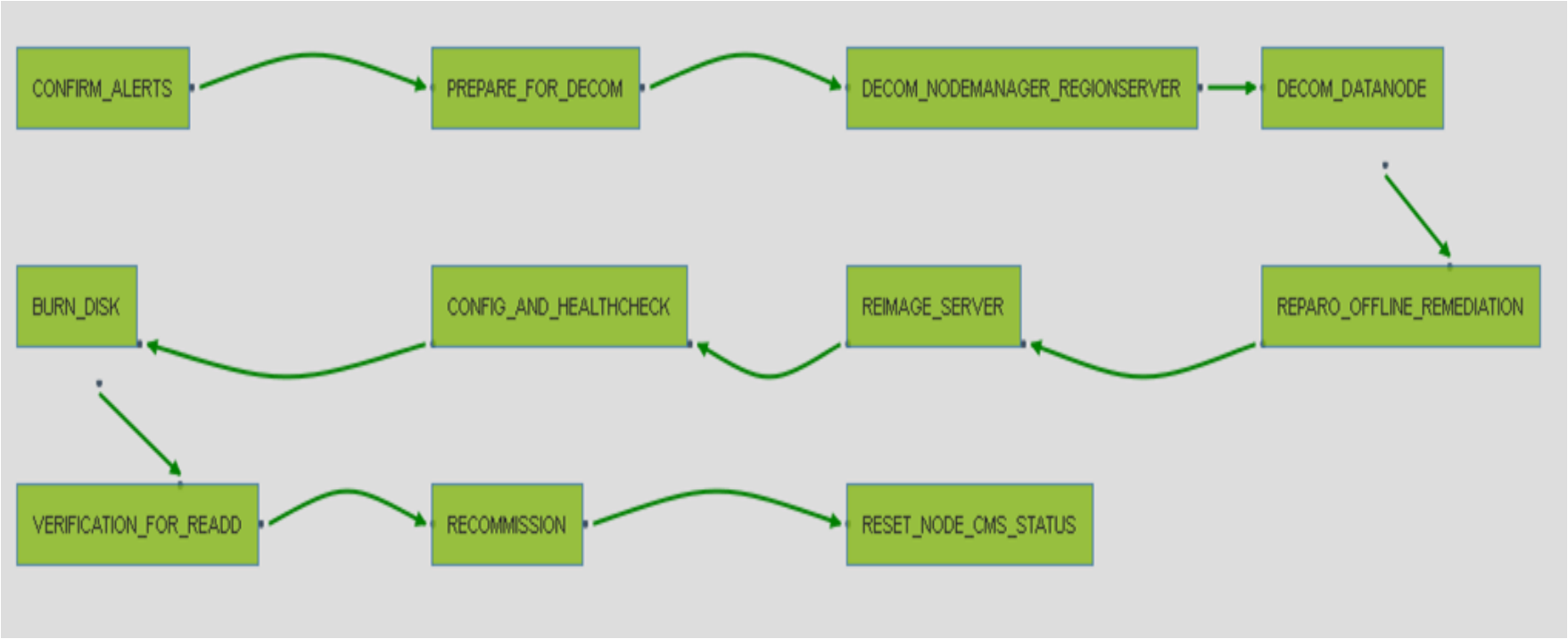
Robot Alert Processing Logic



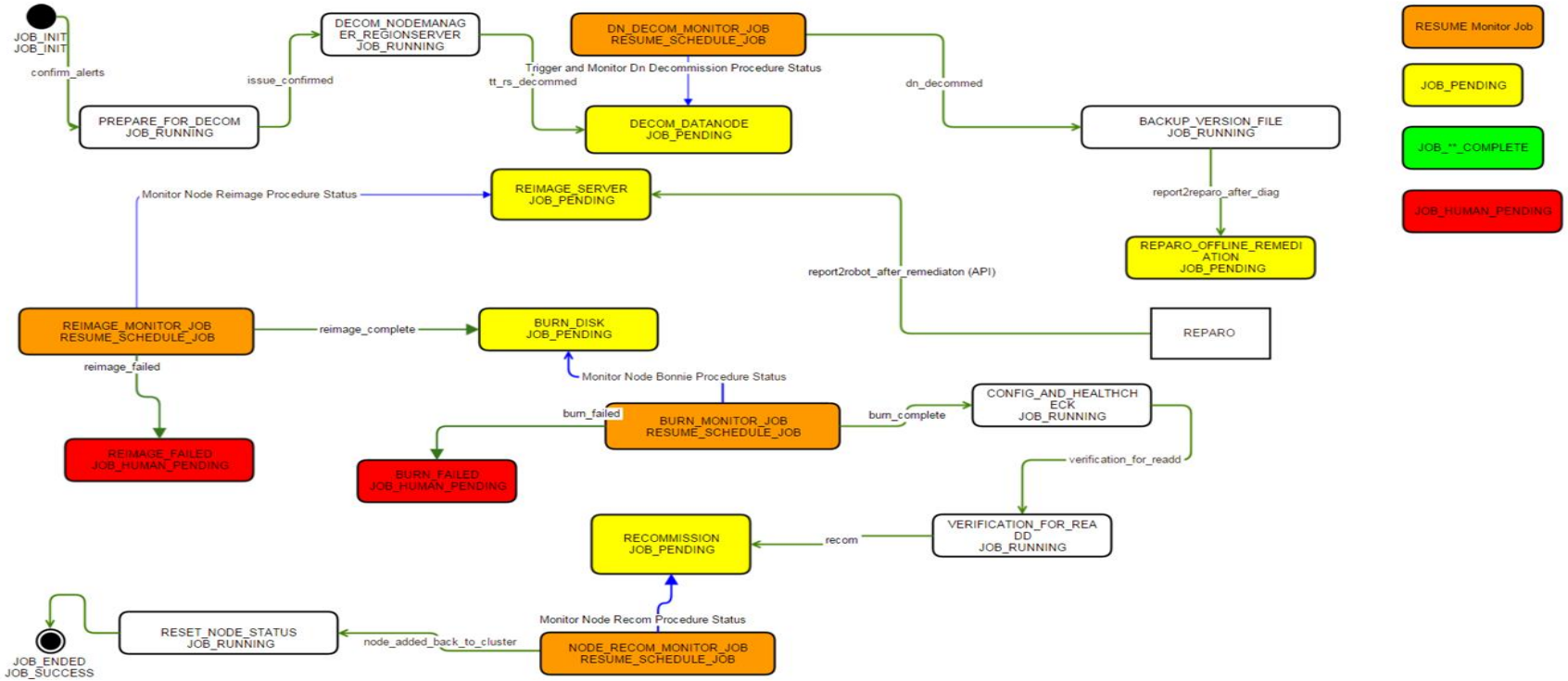
Hardware Maintenance Workflow

- Alert detection
- Node decommission
- Vendor remediation (while node is offline)
- Node OS reprovisioning
- Hadoop Installation && OS configuration
- Node restart
- Burn-in test
- Node health verification
- Node recommission

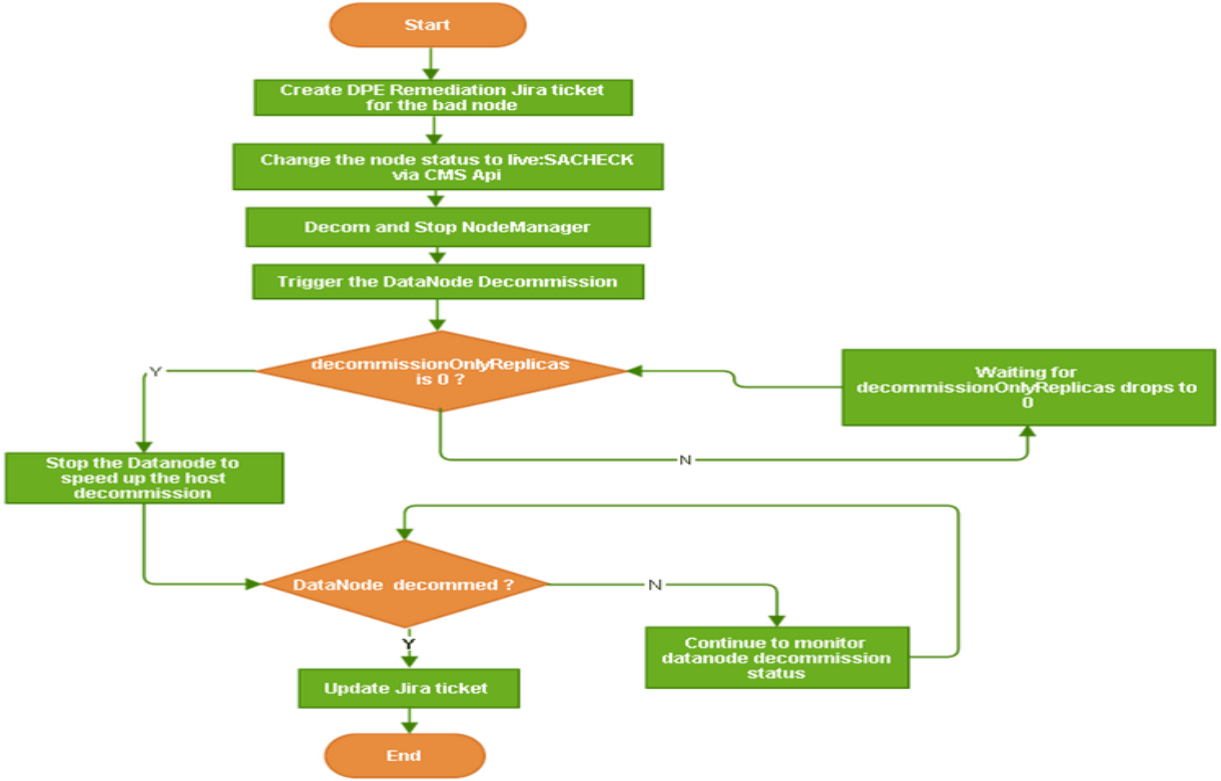
Alerts Drive Workflow



Hardware Maintenance



Node Decommission

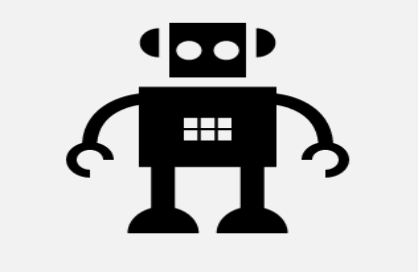


100% Healthy

Vendor

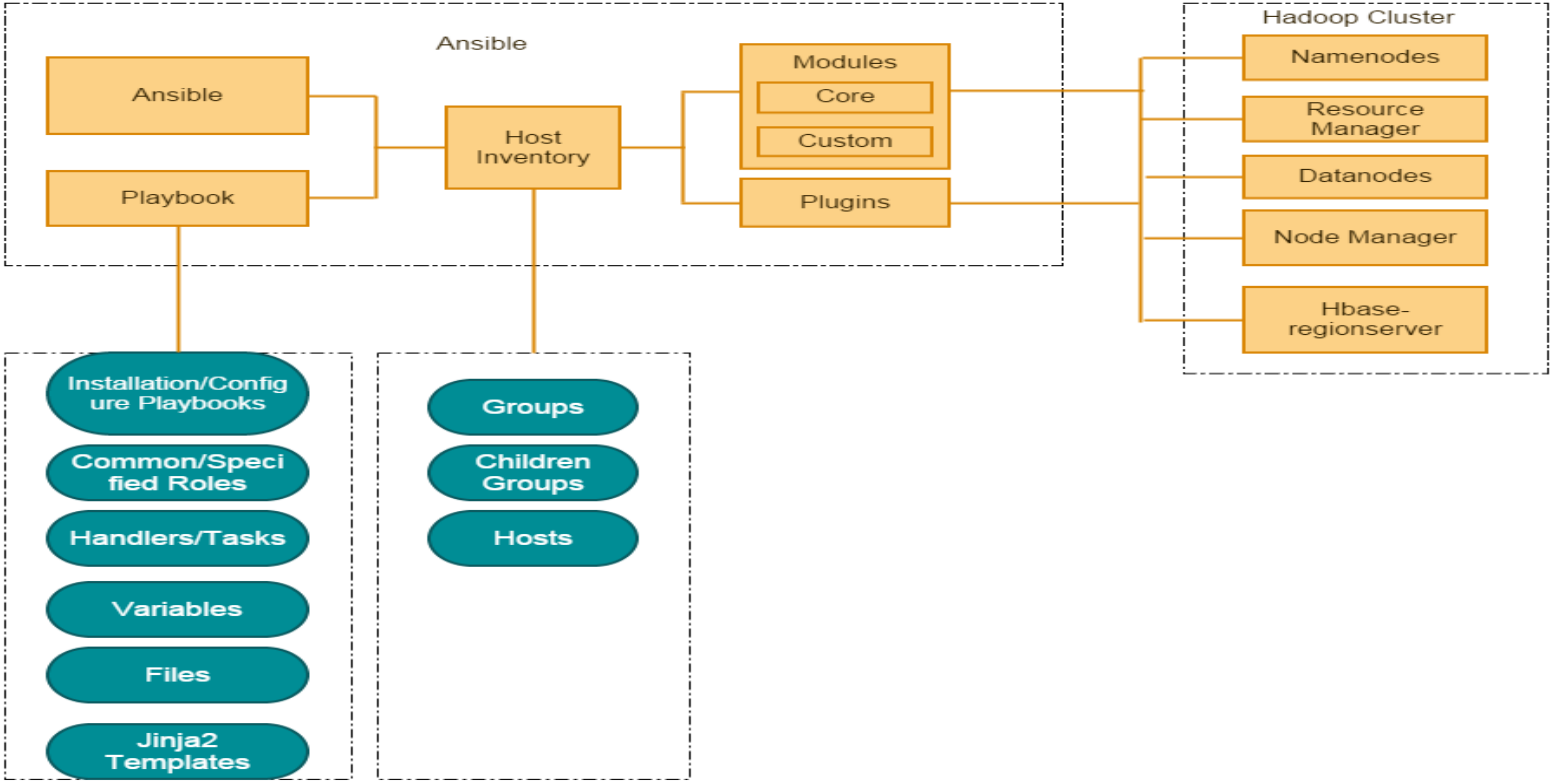
Robot runs remediation job to ensure the Hadoop node is 100% healthy before add it back to the cluster automatically

Vendor returns the server after fixing the hardware



Hadoop Robot

Metadata Setup - ansible



What’s ansible ?

- Configuration management
- Release management
- Automation framework
- Orchestration system
- Distributed batch executor
- No agent
- No server
- Modules in any languages
- Yaml
- Ssh by default
- Strong multi-tier solution

One button hadoop installation and system configuration

We use ansible playbook to install and configure various OS and software packages including Hadoop.

Copy the Hadoop Packages, Configuration and Codes from Source to the Destination Node

↓

Update Symbolic Links to point to the Latest Hadoop Code.

↓

Restart Hadoop Daemon

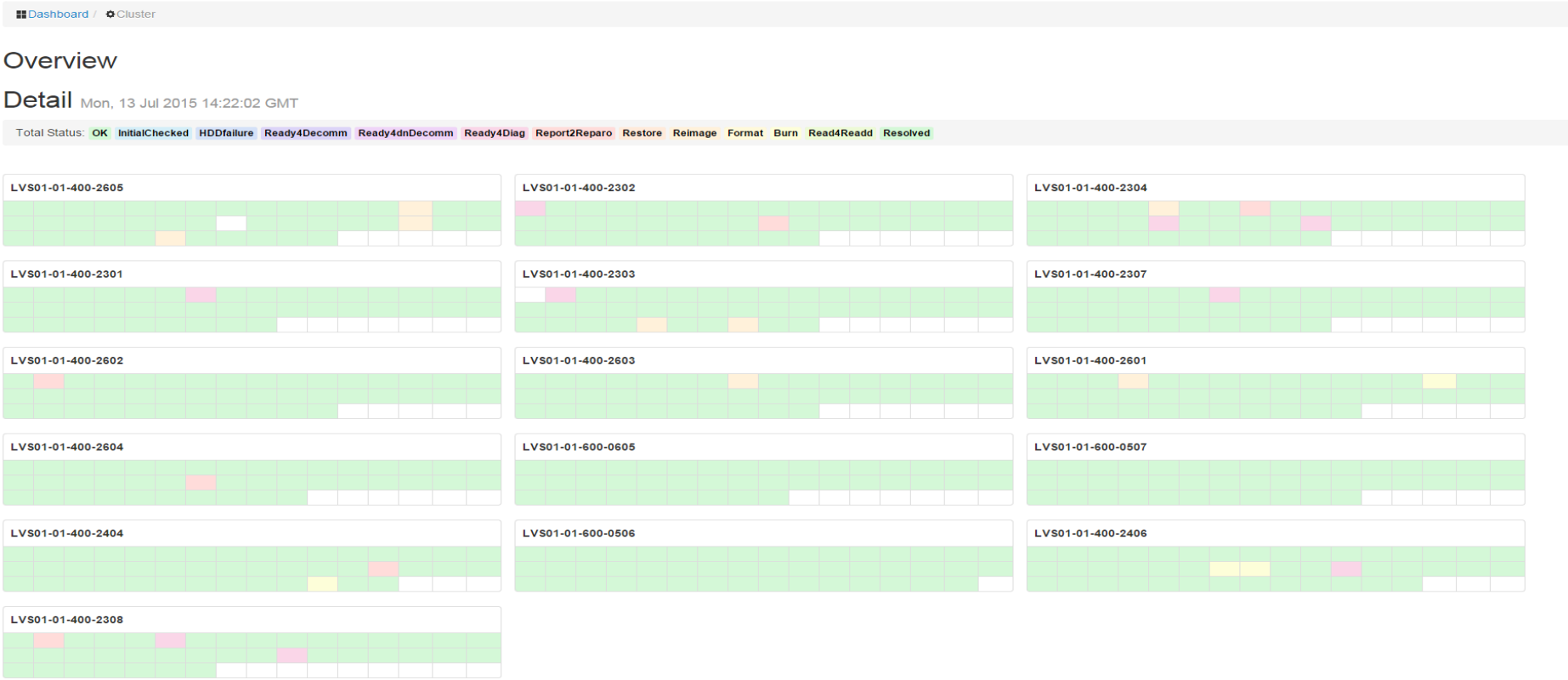
```yaml
- hosts: datanode
  vars_files:
    - vars/main.yml
  gather_facts: yes
  tasks:
     - include: hadoop-config.yml
     - include: enable.yml
     - include: dt.yml

- hosts: masternode
  vars_files:
    - vars/main.yml
  gather_facts: yes
  tasks:
     - include: hadoop-config.yml
     - include: enable.yml
     - include: dt.yml
```

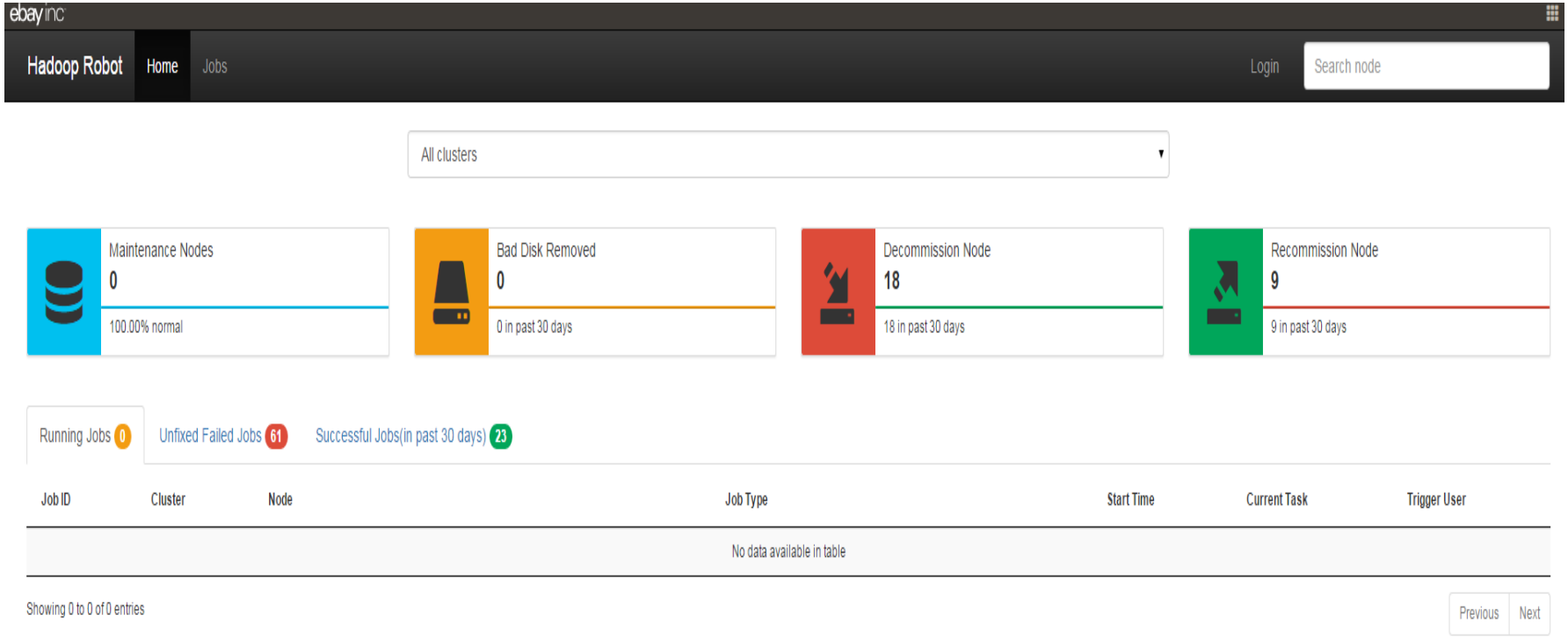
Bonnie++

- We use bonnie++ to carry out a stress-test of the repaired hardware. This not only puts a load on the I/O and disk subsystem, but it also can flush out CPU, RAM and fan/cooling issues.

Cluster level status overview



Unified Hadoop Admin Console



Summit Job

ebay inc

Hadoop Robot | Home | Jobs | Li, Jian | Search node

Automatic Remediation

Initiate Automated Maintenance Workflow - End-to-end repair. Remove a node from participating actively in a cluster, submit it to the reparo hardware repair workflow, reimage, prepare and...

Decommission

Initiate Decommission Only - Remove a node from participating actively in the cluster.

Recommission

Initiate Reimage, Preparation, Burn-in & Recommission Only - Reset a decommissioned node to a fresh state and verify its health, return a node to active participation in the cluster.

Recommission

*Cluster:

*Host name:

*Comment:

*Dry run: Yes

Start | Cancel

Track Robot Job Activities

Hadoop Robot | Home | Jobs | Login | Search node

Home / Node

Search

Job List (past 30 days)

Job ID	Job Type	Start Time	End Time	Status	Result
2344	Decommission	2015-07-09 16:30:42	2015-07-09 16:31:43	Finished	Success
2341	Decommission	2015-07-09 16:11:36	2015-07-09 16:12:38	Finished	Success
2340	Decommission	2015-07-09 16:05:35	2015-07-09 16:06:36	Finished	Success

Job Detail

Labels: Success Running Fail

2015-07-09 16:30:43 -- 2015-07-09 16:30:46: CONFIRM_ALERTS : node fails to pass the initcheck.sh thus the node is really bad
2015-07-09 16:30:46 -- 2015-07-09 16:30:47: PREPARE_FOR_DECOM : prepare for decommission
2015-07-09 16:30:47 -- 2015-07-09 16:30:58: DECOM_NODEMANAGER_REGIONSERVER : decom nodemanager and regionserver on the node
2015-07-09 16:30:58 -- 2015-07-09 16:31:43: DECOM_DATANODE : decom datanode on the node

Robot – Achievement

ebay inc

Hadoop Robot | Home | Jobs | Reports | Login | Search node

Hour saved for each job type:

Cluster: All clusters

Time period: From 2015-12-09 To 2016-01-08 PDT

Generate

- Decommission: 0.5 hour
- Recommission: 2.0 hour
- Automatic Remediation: 3.0 hour
- Remove Bad Disk: 0.08 hour

By Hour Saved | By Job Count