问题：

在云规模系统基础设施中，不同的团队需要频繁地在代码和配置上进行软件更改。但是此类基础设施的规模大和复杂性高，即使在更新组件时出现一个小缺陷，也可能导致广泛的故障，从而对客户造成重大影响。

而且由于测试和生产环境在各个方面存在差异，**准确评估云系统中部署的影响具有挑战性。**

**假阴性：**漏掉一些潜在的问题

**假警报：**误导本地部署运行状况监视器，错误的警报将导致无害的发布停止，并阻止及时应用更改。

解决方案：Gandalf-一种用于安全部署云基础设施的端到端分析服务

Gandalf解决了前面提到的挑战，以确保云基础设施的安全部署。

Gandalf采用自上而下的方法来全面评估部署的影响。

Gandalf连续监控来自基础设施遥测数据的丰富信号集

当检测到系统异常时，Gandalf分析它是否是由部署引起的。如果发现一个不好的推广，Gandalf会阻止它。Gandalf还提供了详细的支持证据和交互式前端。

Gandalf的核心决策逻辑是一个由异常检测、相关分析和影响评估组成的新模型。

设计了一个lambda体系结构的Gandalf系统[6]，将**实时决策引擎**与**批决策引擎**相结合。

具体实现：

实现挑战：

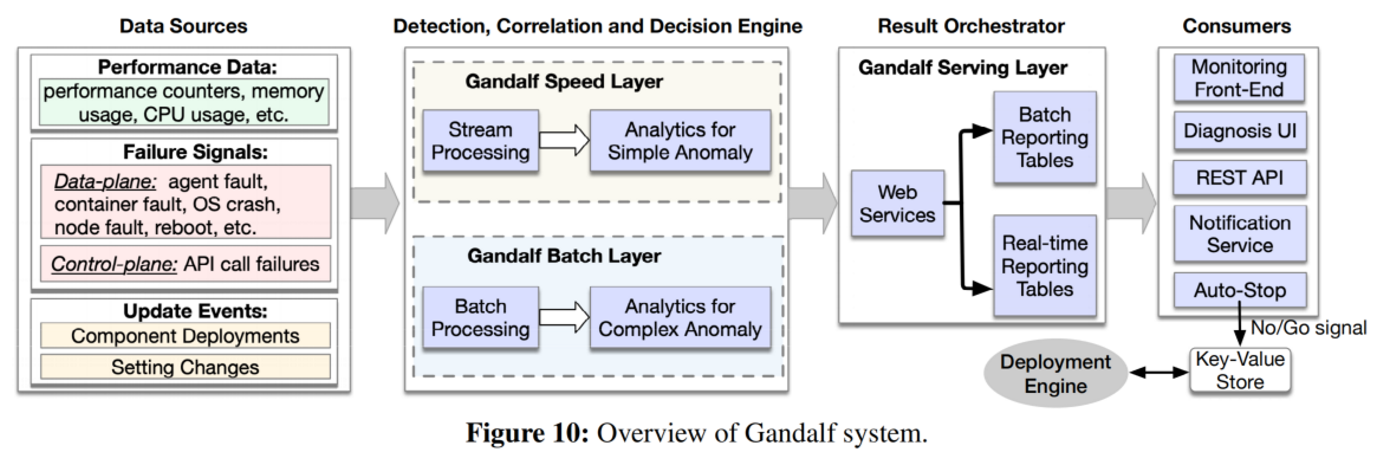
支持系统和信号的变化

处理环境噪音

平衡速度和覆盖范围

识别的罪魁祸首

系统结构



Data Source:

处理三类数据：performance data , failure signals, update events

Gandalf使用来自各种数据源的综合信号，

Stream and Batch Processing:

为了平衡速度和覆盖率，Gandalf采用lambda体系结构[6]，同时具有流式和批处理分析引擎

speed layer使用来自快速管道Microsoft Kusto[7]的数据

Gandalf批处理层使用来自Cosmos的数据

lambda架构允许我们提供快速决策和更高的覆盖率。如果影响范围足够大，可以在窗口内的任何时间触发禁止决策。

Result Orchestration and Actions :

Monitoring and Diagnosis Front-End :

Gandalf提供了一个web前端，为发布经理和开发人员提供了实时的发布监控和问题诊断支持。

算法设计：

相关性模型的步骤：

（1）异常检测从原始遥测数据中检测系统级故障；

（2）相关性分析确定在多个展开中对检测到的故障负责的组件；又分为四个部分，即集合投票、时间相关、空间相关和指数时间衰减。

（3）决策步骤评估受影响的范围，并决定是否应停止部署。

异常检测：

取得什么效果

Gandalf已经在Azure产品上运行了18个多月。

Gandalf系统平均每天处理270K个平台事件，高峰日处理77000个事件，每天在控制平面中记录约6亿个API调用，包括2000多种故障类型。每天分析的总数据量超过20TB。

对于每一个部署，Gandalf可以在5分钟左右的时间内做出决定，将整个生产机队的部署时间缩短了一半以上

Gandalf在进入生产前阻止了99.2%的不良推广。对于数据平面的推出，Gandalf实现了92.4%的准确率和100%的召回率。对于控制平面的展开，Gandalf实现了94.9%的精确度和99.8%的召回率。

优缺点：

云规模系统基础设施中，不同的团队需要频繁地在代码和配置上进行软件更改。但是此类基础设施的规模和复杂性，即使在更新组件时出现一个小缺陷，也可能导致广泛的故障，从而对客户造成重大影响

由于测试和生产环境在集群大小、硬件SKU(库存单位)、OS/库版本、不可预测的工作负载、复杂的组件交互等方面的差异，一些bug可能仍然没有被发现。

准确评估云系统中部署的影响具有挑战性。

假阴性，漏掉一些潜在的问题

假警报，所有这些事件都很容易误导本地部署运行状况监视器，使其错误地将失败归因于无害的部署。这些错误的警报将导致无害的发布停止，并阻止及时应用更改。、

Gandalf

端到端分析服务

解决了前面提到的挑战，以确保云基础设施的安全部署

采用自上而下的方法来全面评估部署的影响。

Gandalf连续监控来自基础设施遥测数据的丰富信号集，包括服务级别日志、性能计数器和进程级事件

当检测到系统异常时，Gandalf会分析它是否是由部署引起的。如果发现一个不好的推广，Gandalf会做出“不去”的决定来阻止它。Gandalf还提供了详细的支持证据和交互式前端，以便工程师轻松理解问题和根本原因。

Gandalf的核心决策逻辑是一个由异常检测、相关分析和影响评估组成的新模型。

我们设计了一个lambda体系结构的Gandalf系统[6]，将实时决策引擎与批决策引擎相结合。实时引擎在部署前后监视一个小时的时间窗口，以检测即时问题；

批处理引擎在较长的时间窗口（30天）中分析系统行为，以检测更复杂的潜在问题。

当Gandalf发现一个不好的部署时，它会自动通知部署引擎停止部署，并向拥有它的团队发出一个带有支持证据的罚单。

部署监控系统

要求：

部署监控系统应能检测到大量基础设施遥测数据中的各种异常-》应将失败归因于所有正在进行的部署中的响应部署，立即停止部署，并为开发人员提供支持证据，以加快进一步的调查

监控系统应准确分析观察到的故障是否由部署不当或其他问题（例如随机硬件故障）引起-》系统不应错误地指责和停止无害部署。

3 Gandalf System Design

3.1 Design Challenges

支持系统和信号的变化

处理环境噪音

平衡速度和覆盖范围

识别的罪魁祸首