사이트신뢰성엔지니어링(SRE)의 요약

Jpub에서 발간한 사이트신뢰성 엔지니어링에 대해 읽고, 자체적으로 해석한 결론입니다. (정확하진 않을수 있습니다. 말 그대로 자체 해석...)

구매는 요기 - http://www.yes24.com/Product/Goods/57979286

SRE란? 구글에서 시스템의 안정성을 증가시키기 위해 활동하는 역활로, Devops보다 한단계 더 발전한 모델이라고 합니다.

SRE역활

1. Site Reliability Engineer 의 약자로, 소프트웨어 엔지니어링과 IT인프라 운영 그 중간쯔음에서 일하는 역활로, in-house tool이나 오픈소스를 활용해 시스템의 안정성과 확장성을 유지하고 개선하는 업무
2. Devops/SRE의 업무 목표는 신속한 서비스 제공을 위해 기업문화, 자동화툴을 이용한 플랫폼 설계/구축 하는 공통적인 업무영역이 있는데, 접근하는 방법이 살짝 다른듯. 어떠한 문제와 개선을 하기 위해서는 "무엇을 해야 하는지?-devops관점"과 "어떻게 할 수 있는지-SRE관점"의 차이점이 있는듯 하다.
3. 기본적으로 기존 운영팀이 처리하던 업무를 이어서 수행하게 되는데, 소프트웨어 엔지니어들이 모인팀이기 때문에 자동화된 소프트웨어를 설계 / 구현하는 팀이다. 다만 전화응대, 수작업, 티켓처리등의 기존 운영팀이 수행하던 업무는 최대 50%정도의 시간만 투입해야 자동화된 시스템 개발할 수 있는 여력을 확보할 수 있게 된다
    1. 운영성 업무를 수행할때는 업무시간에는 최대 2건의 업무만을 담당하게 하고, 업무를 배정받은 엔지니어는 정확 + 신속하게 업무를 수행하고 사후섬토=포스트모텀(Postmorterm)[1)] 을 작성하도록 한다. 2건이 넘어가는 경우 업무의 과중으로 문제점에 대한 관찰력이 저하되는것을 구글은 경험했다
4. 가용성, 응답시간, 성능, 효율성, 변화관리, 모니터링, 장애대응 그리고 수용량 계획에 대한 ownership을 가지게 되는데 이 부분은구체적인 사항은 조금 더 학습이 필요한 영역이다.

기업문화의 변화

1. 포스트모텀 작성시 특정절차에 대해 비난해서는 안되고 실수를 공유하여 스스로 고쳐나갈 수 있도록 유도해야 동일한 문제가 발생시 대응방안들의 절차들이 도출될 수 있다
2. 100% 신뢰성을 가질수 있는 시스템은 없다. 왜냐하면 사용자와 서비스간의 여러가지 요소들(사용자 단말, ISP, 가정의 인터넷, 전력 등)이 있기 때문에 이런 문제들을 해결하기 위해 에러 예산(Error budget)이라는 개념을 도입하였다. (앞으로 기술한 내용중의 에러예산에 대한 내용이 자주 나올예정이다.)
3. 시스템을 구성할때 목표 가용성을 설정해야 한다, 가용성 목표가 정의되면 에러 예산은 1에서 목표 가용성을 뺀 값으로 확보하면된다. 예를들어 목표 가용성이 99.99%라면, 에러예산은 0.01%인 셈이고, 이 에러예산을 초과하지 않는 범위내에서 엔지니어링 업무를 수행하면 된다.

에러예산의 활용방안

1. 구조화된 데이터를 위한 분산 저장소는 응답속도가 빨라야 하고, 높은 신뢰성을 제공해야 한다. 기 시스템을 어떤곳에서는 오프라인 분석을 정기적으로 수행하기 위한 저장소로 사용하기도 하고, 어떤팀은 신뢰성보다는 처리량을 중요하게 생각하는 곳도 있다.
   이런 활용방식을 모두 만족할 수 있는 방법중 하나는 모든 서비스에 엄청나게 높은 신뢰성을 제공할 수 있도록 개발하면 된다. 하지만, 실제로 구현하기에는 엄청난 비용이 발생하기 때문에 현실적으로 어려움이 있다.
2. 개발팀의 업무 목적은 새로운 기능을 출시하여 새로운 사용자를 확보하려고 한다. 때문에 SRE는 새로운 기능을 출시하기 위해 감수해야할 리스크를 활용하는데 사용하는것이 가장 이상적인 활용방안]
3. 목표 가용성 수준을 결정하기 위한 핵심 요소중의 하나가 비용.
    1. 서비스들을 구축하고 운영하는데 가용성 목표 상향시 수익에 긍정적으로 나타나는 효과는 무엇인지
    2. 목표 상향 후 예상되는 추가 수익이 투입된 비용을 상쇄시킬수 있는지. 예를들어 아래 수준으로 반영 할 수 있다.
        - 가용성 상향 목표치 : 99.9 % --> 99.99%, 향상되는 가용성은 0.09%
        - 해당 서비스에서 발생하는 수익 : 1,000,000,000원
        - 목표 상향을 위한 비용 : 1,000,000,000원 * (0.09 / 100) = 900,000원
        - 이 경우 가용성 수준을 0.09% 향상 시키는 비용이 90만원 이하라면 투자 가치가 있으나 90만원을 초과하는 경우라면 예상수익을 초과하기 때문에 목표 상향 가치를 재 검토해야 필요함

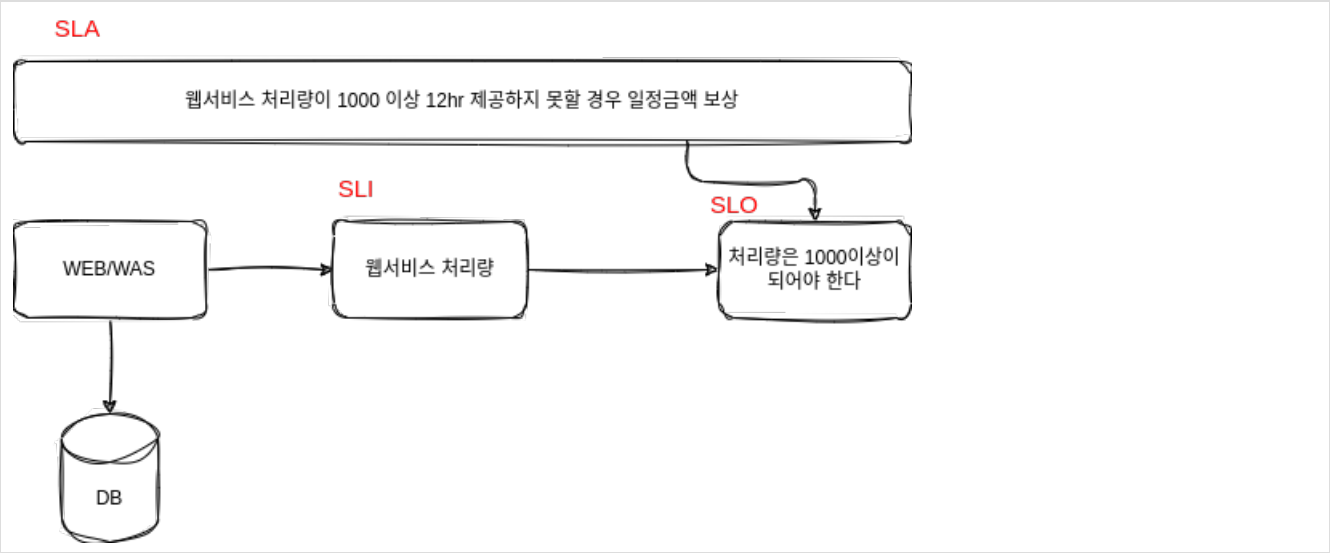
업무 내역

1. 모니터링
    1. 서비스의 소유자가 시스템의 상태와 가용성을 점검하고, 모니터링 전략이야기 말로 철저한 계획하에 수립되어야 하는 업무영역이다. 현재 대부분은 특정 임계치를 초과하거나 상태의 변화가 감지되면 noti하는 방식으로 처리하고 있는데, 이런 방법은 효과적인 해결방안이 아니다. 사람이 개입해서 판단하고 결정해야 하는 절차 자체가 문제가 있는 것이다.

2. 신뢰성이란 문제가 발생하기전 동작했던 평균시간(MTTF)과 평균 수리시간(MTTR)을 의미한다. 이중에 긴급대응의 효율성을 나타내는 수치는 MTTR이 된다.
   사용자가 개입하여 장애를 처리하는것이 시스템이 직접 처리하는 경우보다 MTTR이 3배 이상 증가하는것으로 구글은 검증하였고 실력이 있는 엔지니어가 처리하는 절차도 뛰어나나 잘 만들어진 장애 대응 문서로 훈련된 엔지니어가 보다 더 수행절차가 더 뛰어났다. 때문에 엔지니어들의 장애상황에 대응할 수 있도록 훈련을 지속하고 있다.
2. 변화관리
   1. 구글 경험상 70%가량의 장애는 서비스 중인 시스템의 변화로 인한 문제로 제품의 단계적 출시, 문제발생을 빠르고 정확하게 도출하고 이전버전으로 롤백하는 절차들을 수행하면 장애상황의 최소화가 가능하다.
3. 수요예측과 계획수립
   1. 예측과 계획은 정확하지 않을수 있는 지표때문인지 상당수 수용력을 확보하기 위한 과정을 준비하지 않고 있다.
   2. 수용계획은 자연적 성장(사용자가 활용하면서 생기는 성장)과 인위적 성장(새로운 기능 추가) 모두 고려해야 한다,.

업무지표

1. 안정성과 확장성을 수행하기 위해서는 기준되는 지표가 필요한데, 여기서 나오는 용어가 SLA / SLI / SLO 3가지 용어
   1. SLA(Service Level Agreement)서비스 수준 협약 - 운영팀과 고객간의 서비스 수준에 대해 품질, 가용성 등 구체적인 기준을 설정하는 지표 (구글의 경우 법적인 효력이 있기 때문에 쉽게 변경해서는 안됨)
   2. SLI(Service Level Indicator), 서비스 수준 척도 - 서비스 수준을 측정하는 지표. 예를들어 대기시간, 가용성, 처리량 등의 자료가 포함
      1. SLI의 표준화 (일반적인 정의를 표준화 하기를 권장한다)
         1. 수집간격 : 평균 1분,
         2. 수집 범위 : 클러스터에서 수행되는 모든 tasks
         3. 측정빈도 : 매10초
         4. 집계에 포함할 요청들 : 블랙박스 모니터링으로 수집한 HTTP(GET요청들)
         5. 데이터 수집방식 : 모니터링시스템에 의해 서버에서 수집
         6. 데이터 액세스 응답시간 : 마지막 바이트가 전송된 시간
   3. SLO(Service Level Objective), 서비스 수준 목표 - SLI에서 도출된 지표를 어느정도의 수준으로 품질을 정할것인지 정하는 기준, 발생 가능한 위험과 실행가능성에 대해 조언할 수 있어야 하며, 구글에서는 몇가지 도출안을 제시하였다.
      1. 목표치 설정 가이드
         1. **현재 성능을 기준으로 목표치를 설정하지 말것** : 시스템의 장점과 한계를 이해하고 있어야 하며, 높은 시스템의 이해도를 바탕으로 목표치를 설정해야 하는데 이를 고려하지 않을 경우 시스템에 엄청난 노력을 투입하고, 재설계 없이는 시스템 향상이 불가능하게 될 수도 있게 된다
         2. **최대한 단순하게 생각할것** : SLI를 복잡하게 집계하면 시스템 성능 변화를 명확하게 반영하지 못하고 그 원인 파악이 어렵게 된다.
         3. **자기 만족에 얽매이지 말것** : 응답시간의 저하없이 시스템 부하를 확장하는 것은 매력적인데 사실상 이런 요구는 현실성이 없다.그런 시스템을 설계하게 되면 매우 긴 기간을 추가로 투입되어야 하며 운영비용도 많이 발생하게 된다.
         4. **가능한 적은 수의 SLO를 설정할것** : 시스템의 특성을 확인할 수 있는 최소한의 SLO를 선택해야 한다.
         5. **처음부터 완벽하게 설정은 불가능하다** : SLO 정의와 목표는 시간이 지남에 따라 시스템의 동작을 살피면서 언제든지 다시 정의할 수 있어야 한다. 처음부터 지나치게 높은 목표를 설정하고 이 목표치를 완화하는것보다 처음부터 조금은 느슨한 목표를 설정한 이후에 조금씩 강화하는것이 낫다.
      2. 기대치 설정 가이드
         1. 안전 제한선을 지킬것 : 대고객 SLO보다는 좀 더 보수적인 값으로 설정된 SLO를 지키면 만성적인 문제들이 외부로 노출되기 전 적절하게 대응할 수 있는 여력이 필요하다
         2. 지나친 목표 설정하지 말것 : 서비스층 확보는 실제 사용자들이 경험한 것에 의해 이루어진다.
2. 대충이렇게 정리할 수 있을듯 하다.



위험 요소 관리

1. 기회비용 : 위험에 대비하기 위한 시스템이나 기능을 구현하기 위해 투입할때 발생하는 비용을 의미하며 엔지니어들은 사용자를 위한 새로운 기능과 제품 개발업무 수행이 불가능하다.
2. 서비스중 발생 가능한 위험 요소를 찾아내는 활동을 수행하는데 필요 이상의 신뢰성을 확보하는 활동을 수행하지 않는다. 즉. 목표치가 99.99% 인 시스템을 그 이상으로 초과달성하려고 노력을 하나 넘치게 초과하는 활동은 하지 않는다. 넘치게 초과하기 위한 활동이 이루어질 경우 시스템에 새로운 기능을 추가하거나 운영 비용을 줄기이 위한 기회를 잃어버리기 때문.
3. 서비스 가용성을 판단하는 가장 직관적인 방안은 의도치 않은 장애로 인한 다운타임을 측정하는데, 시간을 기준으로 한 가용성 공식은 다음과 같다. (계획된 서비스 다운타임은 장애라고 판단하지 않는다)

   $$\text{가용성} = \frac{\text{업타임}}{\text{(업타입+다운타임)}}$$

   이 공식을 계산하면 99.99%인 경우 연간 가능한 다운타임은 52.56분 이다. 다만, 멀티AZ와 같이 전세계로 분산된 서비스를 운영하는 시스템의 경우 요청성공율에 기초한 가용성을 정의하는것이 효과적이다.

   $$\text{가용성} = \frac{\text{성공한 요청수}}{\text{전체요청수}}$$

   이 공식을 대입하면 하루에 250만개 요청을 청리하는 시스템은 하루에 250개 이내인 경우 99.99%를 달성할 수 있다.
4. 서비스 위험 수용도를 결정하기 위한 요소
   1. 어느정도 수준의 위험 수용도가 요구되는지
   2. 장애 종유에 따라 서비스에 미치는 영향이 달라지는지
   3. 지속적으로 발생하는 위험 중 어느지점에 서비스 비용을 투입할것인지
   4. 중요하게 고려해야할 서비스 지표는 어떤것들이 있는지
5. 구글에서 정의한 목표 가용성 수준 정의
   1. 사용자는 어느정도 수준의 서비스를 기대하는지
   2. 수익과 직적접인 연관이 있는지
   3. 유료서비스인지, 무료서비스인지
   4. 시장에 경쟁자가 있다면 어느정도 수준의 가용성을 제공하는지
   5. 대상이 개인 고객인지, 기업 고객인지 (기업고고객의 가용성 목표가 조금 더 높다)

Reference

- 사후검토(Postmoterm) - https://scienceon.kisti.re.kr/srch/selectPORSrchReport.do?cn=KAR2005016666

Revision #13
★Created 1 May 2023 04:02:23 by artop0420
Updated 6 November 2023 02:30:36 by artop0420