

QCon 全球软件开发大会 【北京站】2016

Startup大数据平台架构演进之路

App Annie 王佳

QCon

2016.10.20~22

上海·宝华万豪酒店

全球软件开发大会 2016

[上海站]



购票热线: 010-64738142

会务咨询: qcon@cn.infoq.com

赞助咨询: sponsor@cn.infoq.com

议题提交: speakers@cn.infoq.com

在线咨询 (QQ): 1173834688

团 · 购 · 享 · 受 · 更 · 多 · 优 · 惠

7折

优惠 (截至06月21日)
现在报名, 立省2040元/张

自我介绍

王 佳

- App Annie 大数据架构师，大数据团队负责人
- 有12年的软件开发和技术架构经验
- 热爱软件开发行业

联系我：

ramon@appannie.com

概览

- Startup 和大数据，挑战？
- App Annie 大数据架构平台设计原则
- App Annie 大数据架构平台演化

Startup 和 大数据

创业公司构建大数据解决方案时面临的挑战？

Startup 公司

是**勇于探索**可重复和可扩展性商业模式的一家公司，一个合作伙伴或者暂时成立的组织。 – 维基百科

- 在高度不确定的情形下勇于探索和创新
- 快速发展，一直在探索可规模化和可扩展性模式

需求时刻在变化



Startup使用大数据面临的挑战

- 人员成本
 - 大数据平台需要专业技术团队 - 大数据工程师，运维工程师
- 时间成本
 - 组建技术团队需要时间
 - 完整的大数据平台构建需要时间
- 运维成本
 - 保持，维护大数据平台费用昂贵
- 创业公司需求不断变化
 - 如何设计可扩展，可复用的大数据架构平台以满足不断变化的产品需求

It's never too soon for small business to start thinking big.



App Annie 大数据架构平台 设计原则

我们如何应对挑战？



We help build better app businesses

App Annie delivers data and insights to succeed in the app economy















































450名员工遍布全球15个办公室，为客户提供鼎力支持



GREYCROFT IDG Capital Partners IVP GREENSPRING ASSOCIATES
SEQUOIA CAPITAL e.ventures Infinity Venture Partners

首屈一指的应用商店分析和市场数据

游戏	        
社交网络	       
投资	    
平台	      
媒体娱乐	       
其它	       <i>And many more...</i>

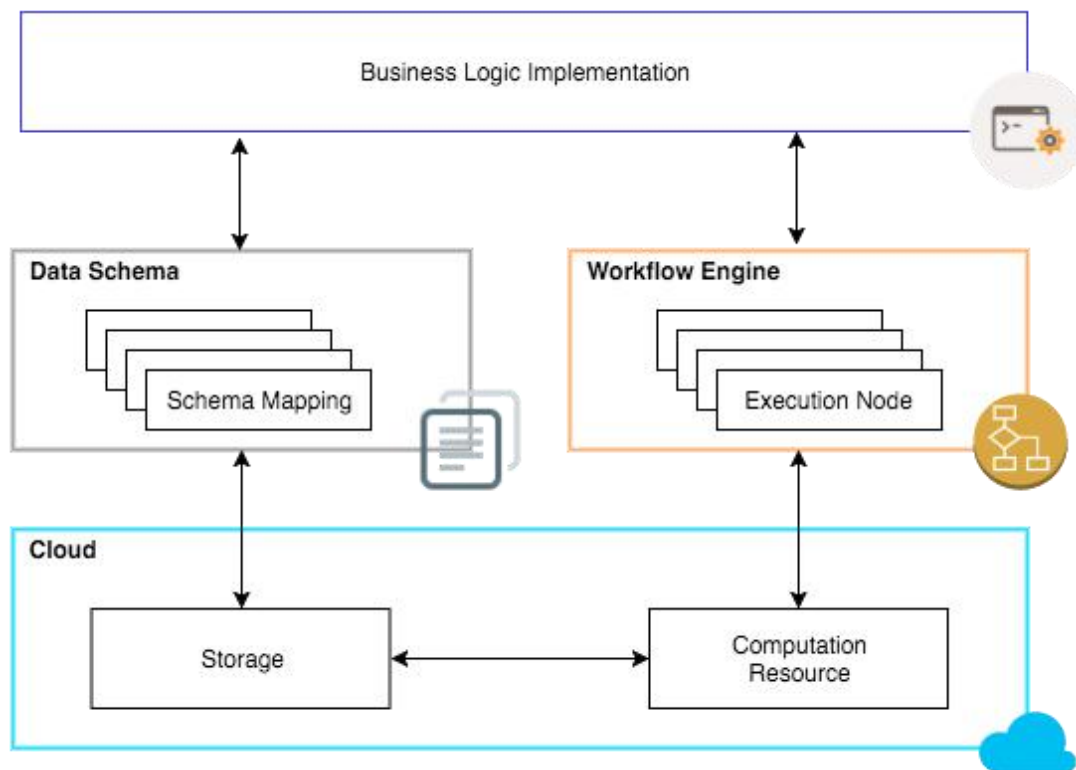
排名前100的发行商中，94%选择 App Annie 产品，拥有超过50万名注册专业用户

App Annie 大数据平台

每天

- 处理 20TB 压缩数据
- 运行 6+ 集群
- 管理 100至200+ 服务器
- 执行 500+ 数据处理任务

设计原则



- 基于云计算服务
- 数据驱动，快速响应业务需求变化
- 使用工作流引擎(Workflow Engine)管理“凌乱”的数据处理任务

基于云计算服务

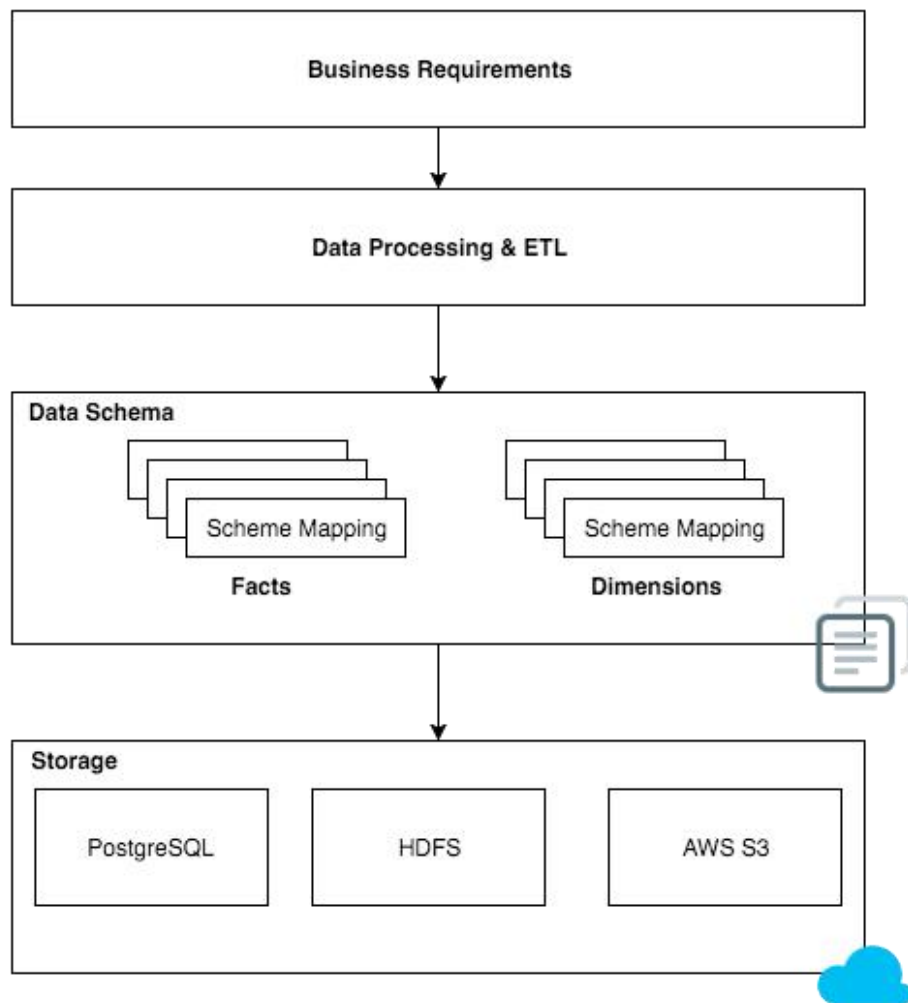
Big Data-As-A-Service

- 超大规模，按需服务
- 高扩展性，高可靠性
- 使用方便，及其廉价

从简单开始...



数据驱动



- 根据业务需求定义 Schema
- 使用 Schema 隔离数据处理逻辑和数据存储方案
- 高度复用 ETL 代码

Schema 例子

```
1  {
2      "type": "s3",
3      "table": "fact/sample",
4      "bucket": "S3_BUCKET_NAME",
5      "access_key": "AWS_ACCESS_KEY_ID",
6      "secret_key": "AWS_SECRET_ACCESS_KEY",
7      "schema": [
8          "id:chararray",
9          "date:chararray",
10         "age_from:long",
11         "age_to:long",
12         "male:chararray",
13         "country:chararray",
14         "description:chararray",
15         "subid:chararray",
16         "type:chararray"
17     ]
18 }
```

- 根据业务需求定义数据映射
- 数据处理代码不受物理存储实现的限制

可维护的代码

```
sample = LOAD ###SAMPLE_SCHEMA###;
```



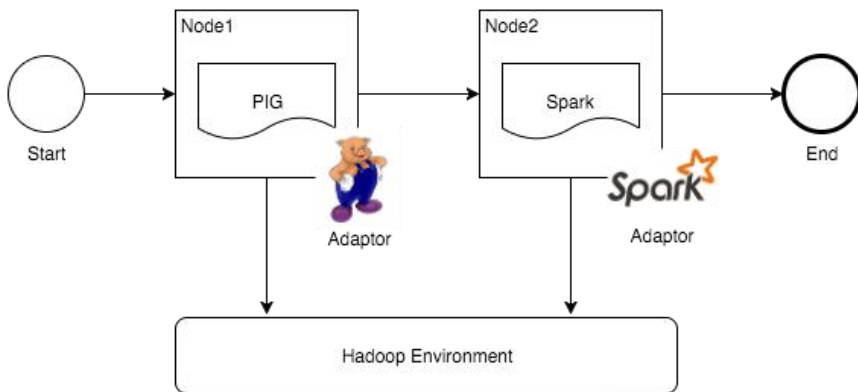
执行前替换

```
sample = LOAD 's3n://S3_BUCKET_NAME/fact/sample'  
  AS (id:chararray, date:chararray, age_from:long,  
      age_to:long, male:chararray, country:chararray,  
      description:chararray, stubid:chararray, type:chararray);
```

- 通过Schema简化Coding，隔离具体的数据存储实现，提高Coding效率
- 集中管理数据结构使代码高度可维护

工作流引擎

- 易于维护
 - 数据处理(Data Pipeline), ETL是一个复杂的过程, 通过工作流引擎可以使各个Job逻辑清晰, 执行状态明确, 实现轻松管理
- 代码简洁
 - 结合Data Schema 可以省略重复的数据定义代码
- 封装与隔离
 - 通过Executor实现不同分布式计算框架的封装



封装分布式计算框架

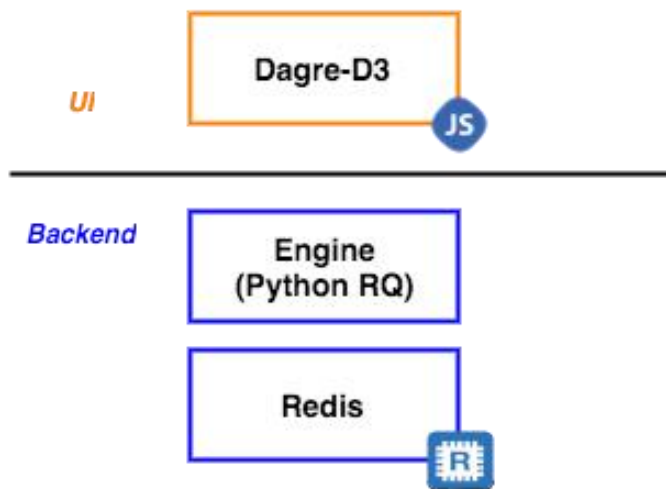


一个简单的Job

工作流引擎选择

自己实现

- 灵活可控
- 功能扩展方便
- 更加贴合业务



开源选择

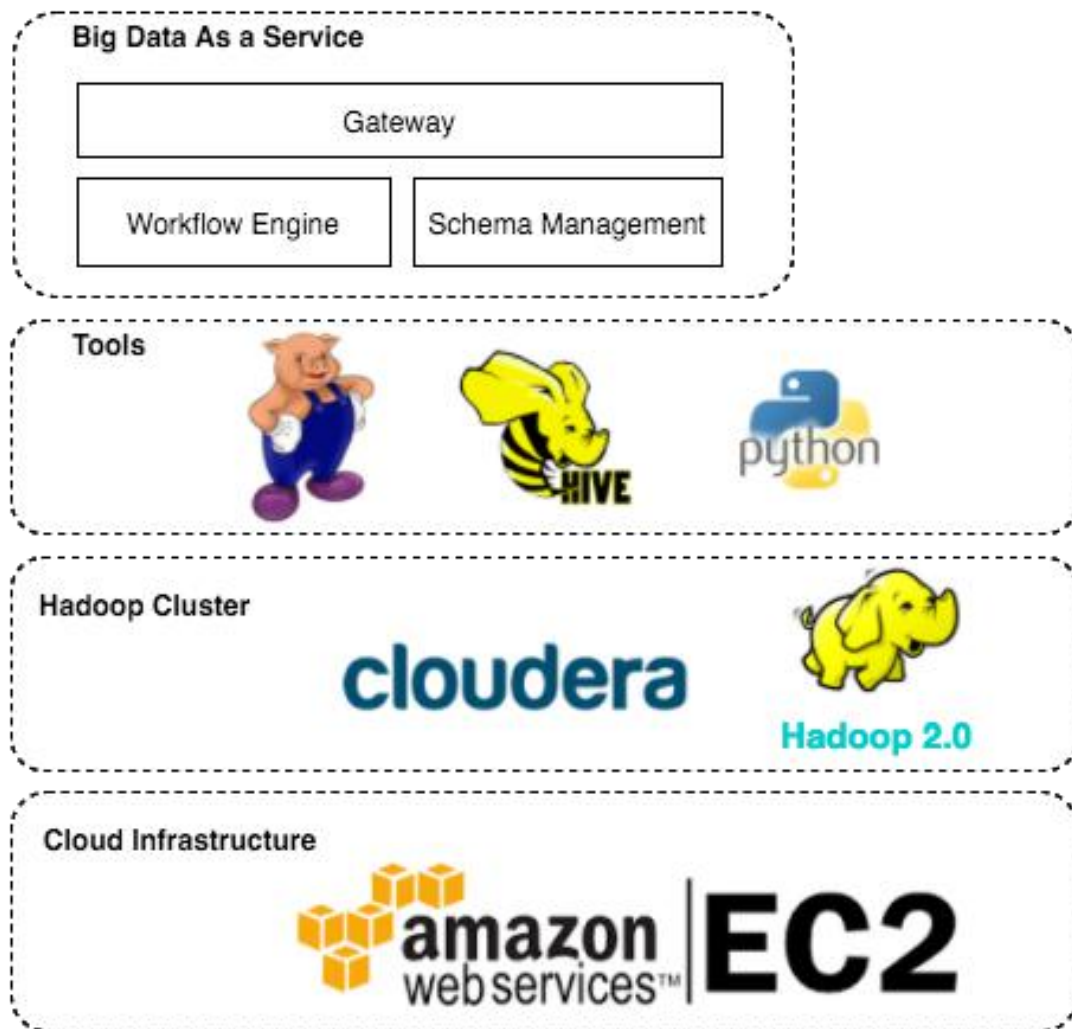
- 社区支持
- 过于抽象
- 不必要的依赖



App Annie 大数据架构平台 演化

从简单开始

从简单开始



- 基于AWS EC2实现资源的弹性伸缩和水平扩展
- 采用CDH简化Hadoop的安装与管理
- 适合规模稳定的小型计算和小规模的数据存储，日处理50GB数据

开发工具和任务提交

- Pig用于构建数据管道(Data Pipeline), ETL和算法模型的实现
- Hive用于支持数据分析和临时性的查询
- Python是我们的主要开发语言, 用于实现各种ETL, 算法模型以及Pig的UDF函数
- 通过Gateway访问集群, 进行任务提交, Gateway是集群的唯一入口

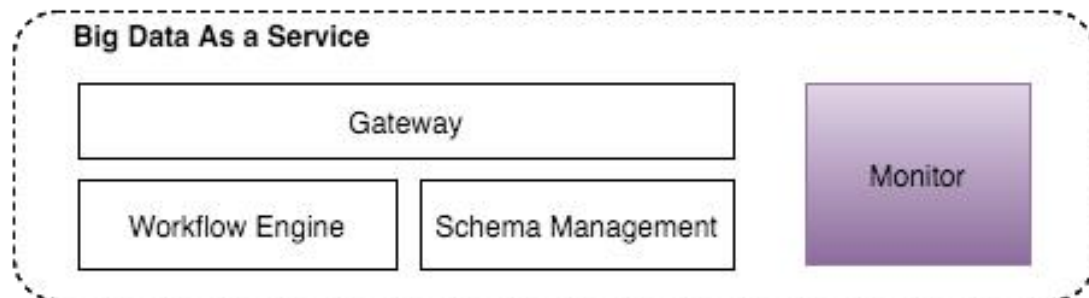
问题

随着数据量的增长...

- 存储、计算资源、集群管理的成本
- 不能灵活地扩充存储资源
- 不能灵活地伸缩计算资源，计算资源和存储资源关联紧密，共享生命周期



演化

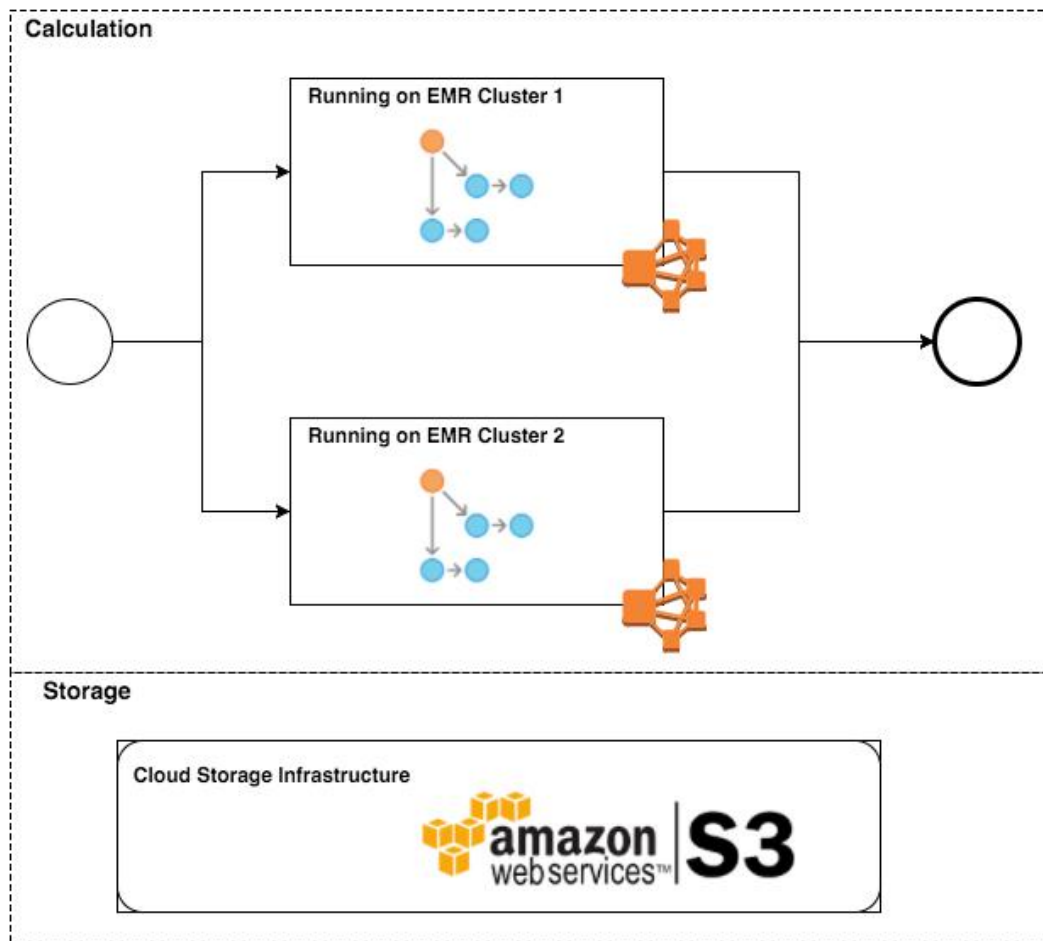


- 分离计算资源和存储资源
- 使用AWS S3获得“无限”的数据存储空间
- 使用AWS EMR灵活管理多个计算集群
- 简化运维工作
- 日处理数据20T至50T

S3作为云存储

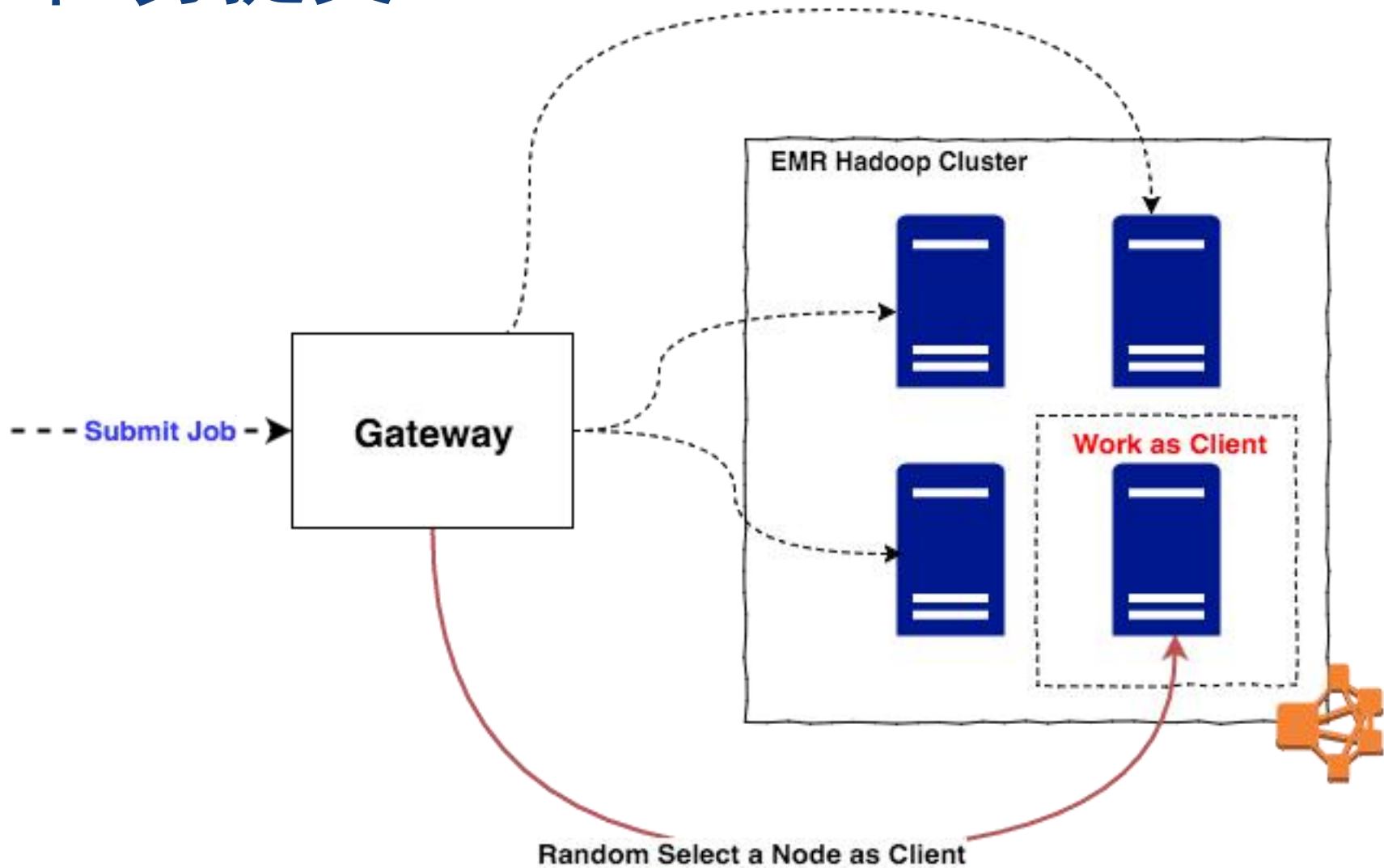
- 高耐久性，高可用性
- 提供版本机制，避免数据误删操作
- 可伸缩的存储，按需使用付费
- 提供接近“无限”的存储空间
- 允许多个集群同时读写同一个 S3 buckets

灵活的计算资源配置

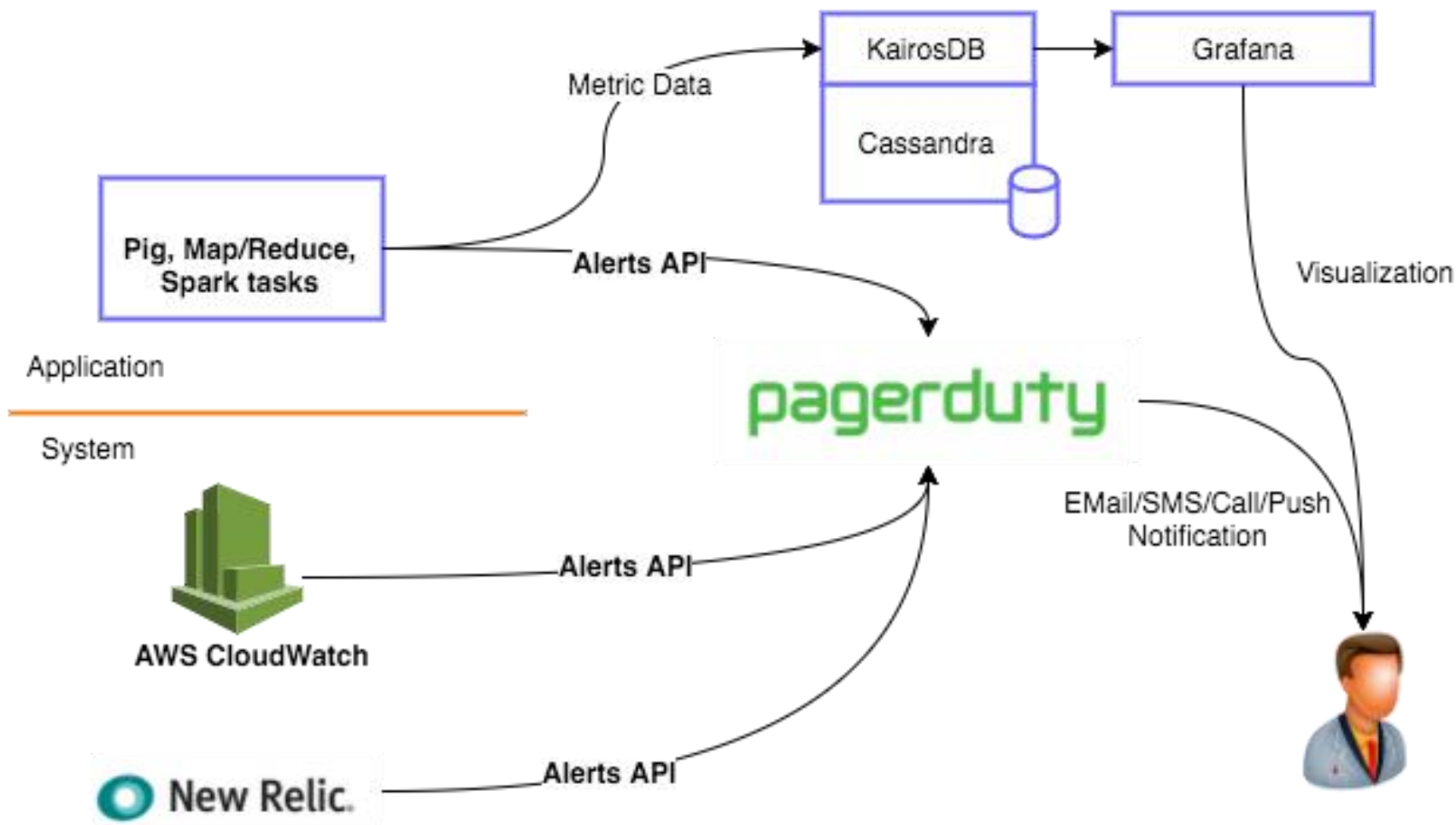


- 多AWS EMR集群 + S3
- 灵活配置HDP集群满足不同的需求，单个集群规模从2到200+
- 通过Workflow灵活配置集群，按需使用计算资源
- 计算资源与存储资源有不同的生命周期

任务提交2.0



监控



问题

随着业务快速增长，需要处理多个国家、地区的数据...

- 单数据中心导致原始数据必须从远程导入，增加数据传输逻辑的复杂性和成本
- 增加数据处理逻辑的复杂度
- 单点失败问题

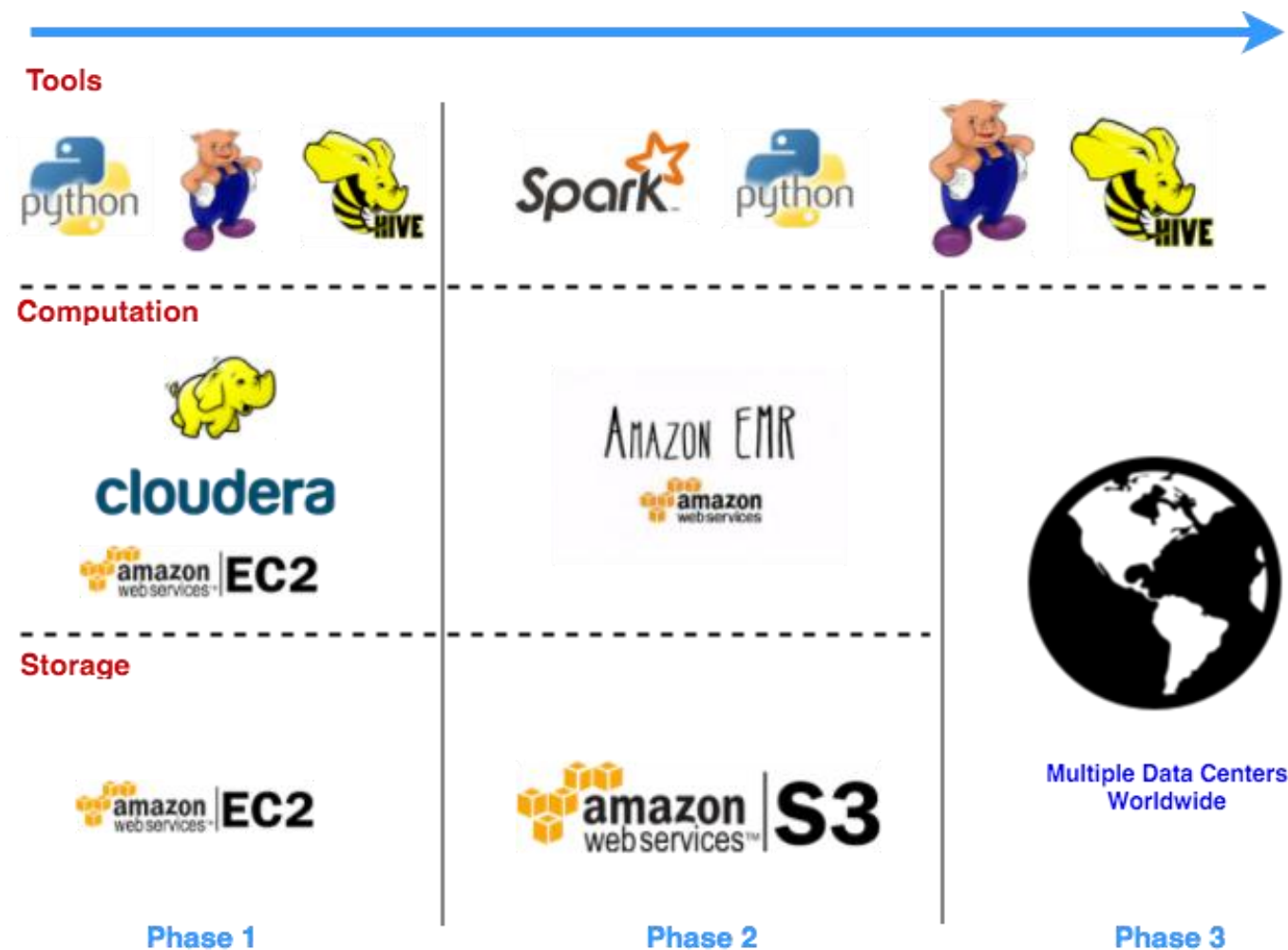


多数据中心



- 通过AWS的云服务完成数据中心的多区域托管
- 数据就近处理，架构简单明了
- 部署简单，易于维护

技术演化 总结



We Are Hiring

热招职位

- Big Data Engineer
- Full Stack Engineer
- Chief Architect
- Senior DevOps Engineer

请将简历投递至：
recruiting-apac@appannie.com

App Annie 为员工提供

- 尽可能升级您的工作装备：



- 尽可能改变您的生活状态：



- 尽可能降低您的生活成本：



Geekbang
极客邦科技

InfoQ

THANKS!

App Annie

International Software Development Conference