面向多云环境的数禾IT平台架构探讨

SA. 尹广东

Date: 2020-03

议题

一、回顾过去（单一云平台）

I. 数禾+AWS共同成长历程
II. 最近一年AWS BJS平台服务改进措施

二、审视当下（正在走向多云）

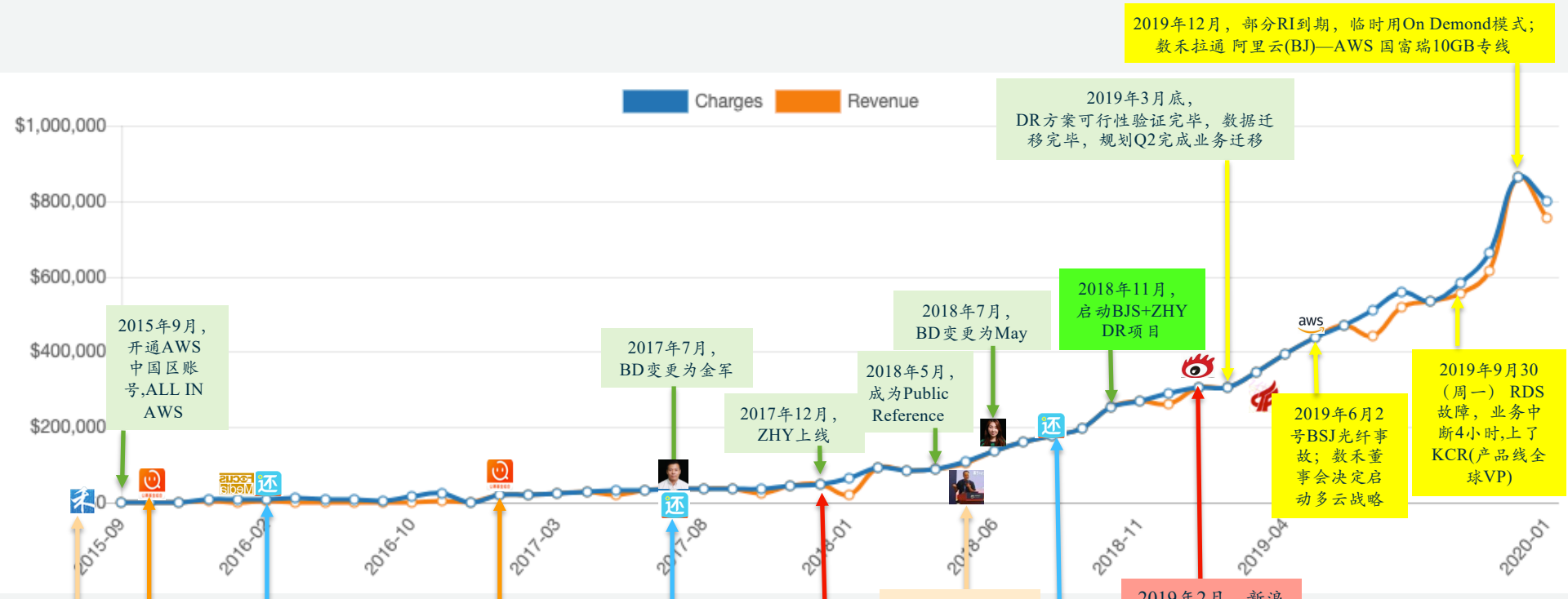
I. 2020数禾主要项目保障计划
II. 大数据平台@AWS的使用状态汇报

三、展望未来（用好多云）

I. 数禾未来对云平台的使用规划
II. 数禾对业务连续性的规划
III. AWS BC Lab（ Business Continuity ）方案介绍
IV. 多云方案的几种场景

1.1、数禾+AWS携手相伴共同成长历程回顾

2020年2月，数禾申请再拉一根10GB专线
阿里云(BJ)—AWS 国富瑞 做专线HA



由西云数据运营
由光环新网运营

1.2、最近一年AWS BJS平台服务改进措施

1. 2019.6.2光纤故障后，AWS BJS进行了如下调整：
   ① 距离靠近的线缆重新部署
   ② 增加新光缆，并且使用不同的地下管道（不同的线路供应商）
   ③ EC2 API控制平面优化可进行快速切换
   ④ 网络流量从受损的AZ快速切换至健康AZ改进

2. 目前BJS Aurora（基于3 AZ）已经上线；

2.1、2020数禾主要项目保障计划

ID	项目描述	使用服务	业务诉求	完成时间	AWS保障小组
1	远程安全办公桌面项目	WorkSpaces	WorkSpaces实时完成，AWS SA+TAM现场回访排除遗留问题。	2020 Q1	BD、SA、TAM
2	对接腾讯广告投放系统	DynamoDB	项目测试已经完成，正在大规模投入生产	2020 Q2	BD、SA、TAM、AoD（韩思捷）
3	自建K8S迁移至EKS	EKS	数禾核心业务全部是基于EC2上自建的K8S进行的容器化部署；运维工作量繁重，出现故障troubleshooting困难，会影响业务效率；	需和数禾沟通确认	BD、SA、TAM、AoD(尹振宇)
4	BJS Aurora替换MySQL	Aurora	数禾业务系统依然在高速发展，对数据库的极致性能、高可用性要求极高，用Aurora替换MySQL，可以很好满足数禾对数据库的高性能、高稳定性需求。	需和数禾沟通确认	BD、SA、TAM、AoD(邵虎)、CSM
5	ProServe业务连续性	ProServe	数禾已经由早期几十人的初创团队，在两年的时间里迅速发展到500+的人员规模，从企业自身发展的角度，已经到了需要对IT治理的规范性需要进行全面评估及提升的时期。	需和数禾沟通确认	BD、SA、TAM、CSM、ProServe

2.2、大数据平台@AWS的使用状态汇报—1

1）目前数禾大数据系统的使用情况

- 常驻的EMR集群有14个，其中4个是Jupyter的训练集群，剩下10个中一个是Flink的实时计算，另外一台是Presto的EMR，基于CDH的HDFS文件系统，其他都是跑批的（一小时或15分钟准实时类型）。
- S3的流量的读峰值超过500GB/分钟（平均10GB/s），写峰值60GB/分钟。
- S3数据增量的波动较挺大。长期来看，每天的增量TB级别，每天大约1TB多不到2TB的增量。
- EMR中主要使用的组件为Hadoop，Yarn，Spark，Flink，Presto，Hive，Hue，Tez，Sqoop。

2.2、大数据平台@AWS的使用状态汇报—2

2）AWS大数据服务在数禾的使用情况

- AWS大数据服务（EMR）在数禾稳定运行了3年以上，稳定性得到了生产环境的验证；
- 数禾大数据系统大量使用了Spot实例和RI，节约了大量成本；
- AWS EMR和开源社区主版本一致性很高，避免了平台绑定。

3）阿里云大数据相关服务的局限性

- 阿里对开源 Hadoop 有进行较多定制化，而且对社区版本跟进速度不快；
- 当前阿里大数据版本问题较多，阿里 EMR 对 Spot 的支持较弱，无自动扩容收缩功能，也不支持加密，权限颗粒度较粗，也没有 SLA 保障；
- 需要根据客户需求进行改造，数禾在数据处理和分析部分需要同步做适配工作，存在较大业务和系统风险；
- 阿里大数据流量峰值极限单账号12GB/s（AWS 实际有用户S3吞吐达到了260GB/s ）

2.2、大数据平台@AWS的使用状态汇报—3

4）数禾大数据使用后续优化建议

- AWS平台EMR版本稳定，CDH迁移后使用EMRFS完全满足要求并可进一步降低成本，更容易构建数据湖。
- AWS可以提供专家队伍并在CDH集群迁移AWS EMR期间提供驻场专家支持。
- AWS是全球各金融机构(如Capital One)云平台及大数据平台的提供者，在今后各领域可以引入海外专家一起进行经验交流

3.1、关于数禾未来对云平台使用规划的讨论

① 数禾目前启动了应用迁移，下一步对于云平台的使用规划及方向？

② 基于多云场景（包括自建数据中心），数禾主要的考虑点与要求是哪些？

③ 数禾当前执行的应用迁移是否有利于提高业务连续性？

④ 未来架构，如何实现更好的业务连续性、稳定性、避免平台绑定？

3.2、数禾对业务连续性的规划

① 从AWS迁移至阿里是否能够解决业务连续性问题？（AWS是运营时间最长的云厂商，依然会出现各种问题；阿里云作为后来者，从以往历史上来看，出问题的频率比AWS要多很多）

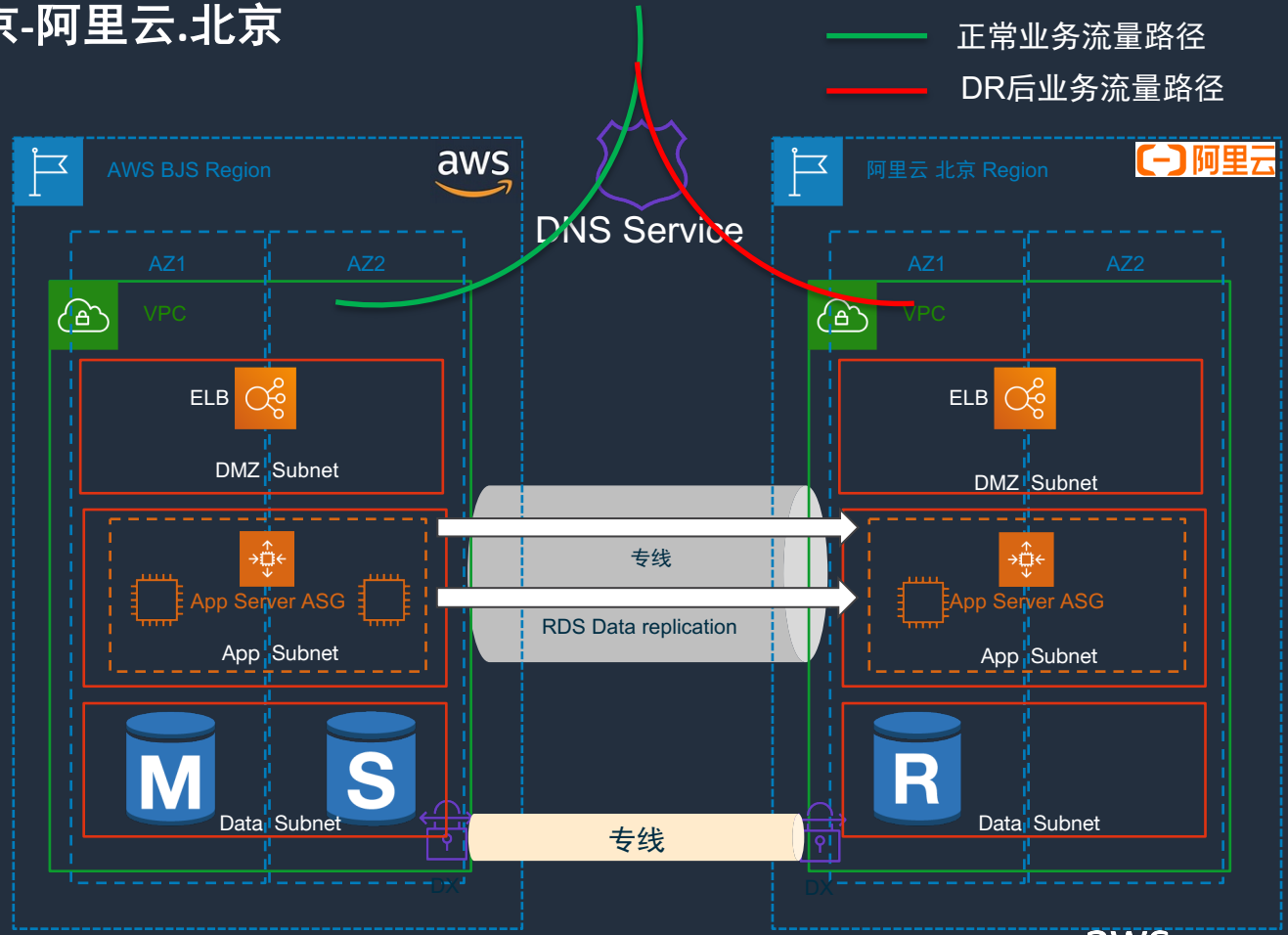
② 多云可以提供更好的保障（两家云平台同时出问题的概率极小），建议数禾拿一个APP做成双活进行测试。AWS愿意投入资源和数禾一起设计方案并测试验证。

多云容灾架构：AWS北京-阿里云.北京

具备完整功能的缩小版（或者完整版）环境始终在灾备区域中运行。

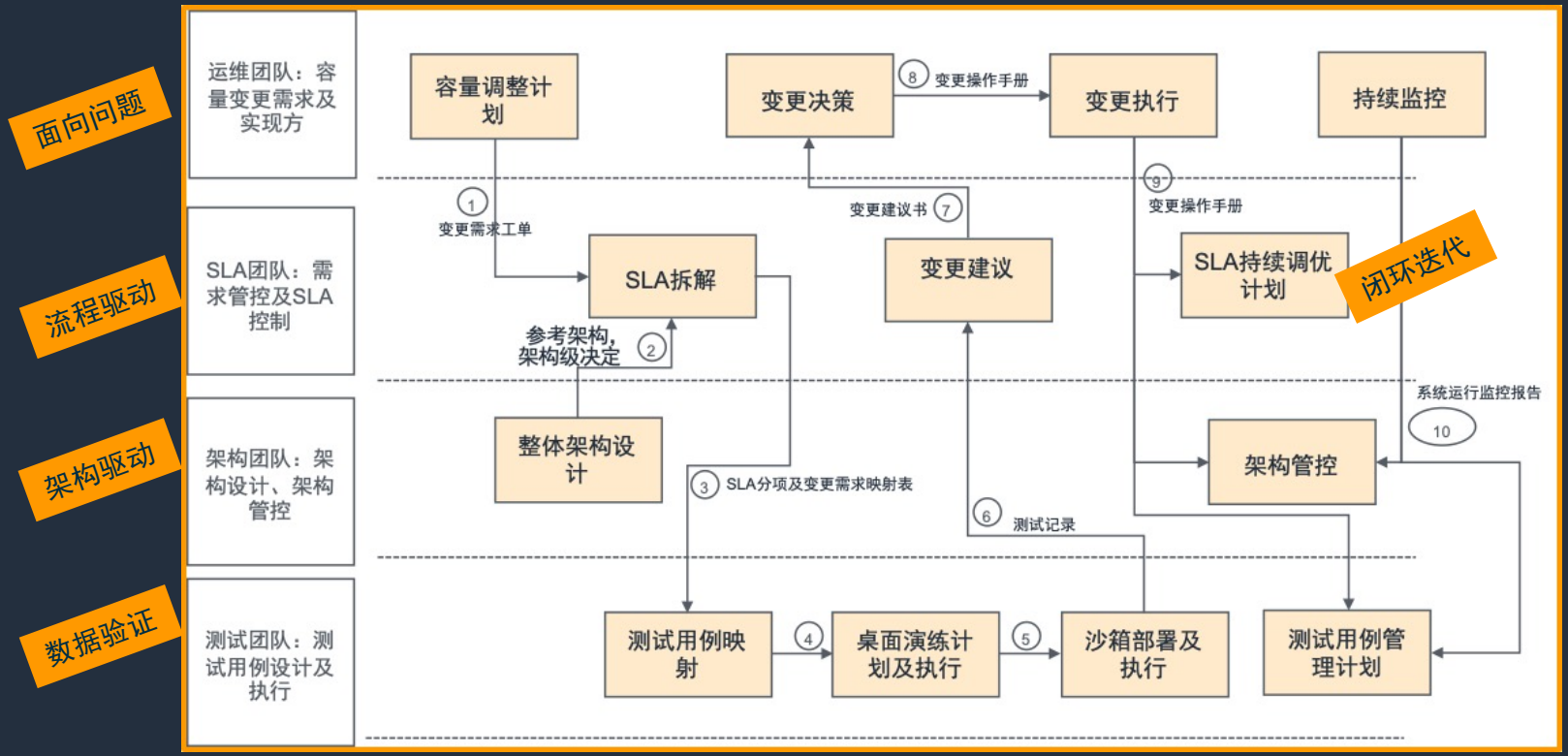
通过DNS服务进行流量切分，将用户访问数据流按照预设比例分配到不同的云平台环境。

假如有一个平台发生故障的情况下，另一环境将按比例放大以用于支撑全部生产负载，并且将更改DNS记录以将所有流量路由到另一个健康区域



AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营

AWS经大量实践、综合思考相关方法论，总结出融合持续集成、持续交付和持续运维的业务连续性工作方法。
AWS CI/CD/CO连续性方法具备五个特点：



在中国某领先金融科技机构实践时，针对变更是最影响业务连续性的特点，选择变更为关键先导场景。

Lab 前

- 问题：因业务需求不确定、流程不清晰、架构不完整和技术不成熟导致的应用服务不稳定，造成业务团队不满意
- 问题分析：
  - ✓ 从故障记录看，客户故障？%是因为变更导致。
  - ✓ 从金融行业的数据看，一般是40%~70%因变更引起。如XX行，？？年是？%是因为变更导致的。
  - ✓ 从互联网企业看，如XX，%是因为变更引起的。
  - ✓ 因此，变更是重要的触发事件，解决问题的方法应围绕变更、基于流程展开。
- 解决问题的方法：以一个具体的变更流程为切入点，经过体系化设计，形成具体的操作手册
  - ✓ 环节：变更需求管理、SLA拆解、架构审核、变更计划制定、变更操作实施、系统验证完成以及变更演练
  - ✓ 团队：业务、开发、运维团队
  - ✓ 技术：应用、中间件、数据库、基础设施
  - ✓ 设备：各供应商（包括AWS与自建机房）
- 预期效果和交付件：
  - ✓ 变更操作手册：当前客户团队能手工操作、操作之后能沉淀数据、未来能转换成自动化Code
  - ✓ 客户团队对变更有敬畏之心
  - ✓ 客户团队对现代化运维体系的方法论有初步认识

Lab 后

Lab 的发现：
- 业务需求：容量需求管理方法不确定、变更需求缺乏管控
- 流程不清晰：变更流程不同环节的衔接不清晰
- 架构不完整：缺乏从上而下、系统性的架构管控、设计、执行和跟踪
- 技术不成熟：测试工具不完整

Lab 的工作内容：
- 环节：端到端完成
- 团队：包括稳定性小组、Datadog开发、运维和测试
- 技术：涉及应用、中间件、数据库、基础设施
- 设备：包含AWS环境

Lab 的工作成果：
- 手册：完成
- 客户团队对变更有敬畏之心：初步
- 客户团队对现代化运维体系的方法论有初步认识：具备

3.4、 多云方案的几种场景

- 方案A：核心容灾+大数据保留AWS（架构最优）
- 方案B：大数据保留
- 方案C：大数据迁走

Thank You!

附录：阿里云OSS的带宽极限

https://help.aliyun.com/document_detail/54464.html?spm=a2c4g.11186623.4.4.75847f36hT5gnu

上传/下载文件	• 通过控制台上传、简单上传、表单上传、追加上传的文件大小不能超过5GB，要上传大小超过5GB的文件必须使用断点续传方式。 • 断点续传方式上传的文件大小不能超过48.8TB。 • 同一账号在同一地域内的上传或下载的带宽缺省阈值为：中国内地各地域10Gbit/s、其他地域5Gbit/s。如达到该阈值，请求的latency会升高。如您的业务（如大数据离线处理等）有更大的带宽需求（如10Gbit/s~100Gbit/s），请联系售后技术支持。 • OSS支持上传同名文件，但会覆盖已有文件。

附录：关于阿里的故障（2012-2015）

2012年10月30日，由于电力故障阿里云部分服务器30余分钟无法正常访问， 事后阿里云为此次因电力故障受影响的用户统一提供百倍赔偿。

2013年1月18日，阿里云机房发生临时故障，部分用户服务器无法访问，20分钟修复。1月23日，阿里云发生网络系统故障，OSS服务无法正常进行，故障持续长达6小时。

2014年11月14日，由于市政施工导致运营商光纤受损，阿里云杭州可用区D网络故障，具体影响ecs、rds、ocs等云服务半小时左右，受此事件影响，当天不少P2P平台网站无法打开。

2015年6月21日，一些使用阿里云香港数据中心的用户发现服务出了问题，服务中止12小时。此后，阿里云公告称由于运营商电力问题造成香港机房故障。

2015年9月1日，有多位用户在微博爆出运行在阿里云上的系统命令及可执行文件被删除，严重影响线上服务及运维。阿里云官方声明称，是由于云盾升级触发bug，导致少量文件被系统误删除。对于受影响的客户，将立即启动百倍时间赔偿，并避免类似失误再次发生。9月3日，阿里云云盾负责人吴翰清撰文阐述事件真相“工程师粗心大意写错一行代码”，并向受影响的用户道歉。

附录：关于阿里故障（2016-2019）

2016年7月6日，阿里云北京机房内网发生故障，导致大量互联网公司业务受到影响。阿里云工作人员表示，10点20分阿里云北京区开始出现故障，接近11点20分恢复正常。

2018 年 6 月 27 日，阿里云出现大规模访问异常，图片服务等产品无法正常使用，官网账号也无法登陆。

随后，阿里云正式发布通告称，于北京时间 2018 年 6 月 27 日 16:21 分左右，阿里云官网的部分管控功能，及 NAS、OSS 等产品的部分功能出现访问异常。阿里工程师正在紧急处理中。

2019年3月2日的华北区，阿里云“华北2地域可用区C部分ECS服务器（云服务器）等实例出现IO HANG（IO无响应，即磁盘无响应）”。那次故障影响较大，震动了社交媒体圈。当时阿里云宕机3小时，包括广告传媒、赛事直播、视频网站、软件服务、云服务器提供商和阿里集团核心业务等都受到了程度不同的影响。

http://www.360doc.com/content/19/0303/18/46043804_818852423.shtml
https://baijiahao.baidu.com/s?id=1633781257633053092&wfr=spider&for=pc

典型情况下，BC Lab涉及的工作计划包括：

时间		内容
第一天上午	09:00-09:30	AWS BC Lab介绍
	09:30-10:30	在技术架构的层面，AWS CI/CD/CO方法论和最佳实践介绍
	10:30-12:00	客户架构review和重点隐患选择，Pre-lab选择的切入点回顾
第一天下午	14:00-17:00	客户团队针对其他应用，参考AWS的示例，形成其他应用的可用性改进建议
	16:00-18:00	AWS领域专家和客户团队通过问题发生的概率、对业务的影响程度和解决的难度来形成改进路线图
第二天上午	09:00-10：00	在架构管控的层面，介绍AWS的方法论和最佳实践
	10：00-11：00	挑选挑选1~2个重点应用，来讨论需求的提出和SLA的定义
	11：00-12:00	挑选挑选1~2个重点应用，来讨论架构的决定、实施和运维模式
第二天下午	14:00-17:00	客户团队针对其他应用，参考AWS的示例，形成其他应用的建议
	17:00-18:00	AWS结合客户团队的管控实践、讨论形成从易到难的实施路线图（例如方法、工具和需要的能力）
第三天上午	09:00-10：00	在运维流程的层面，理解AWS针对金融客户的方法论和最佳实践
	10：00-12：00	结合客户之前遇到的可用性问题，针对运维中事件、问题、变更和配置的管理，挑选一个流程（例如变更流程），端到端（包括业务、应用、技术架构、AWS服务等）的讨论如果建议一个可实施的流程（流程、人员、管控、工具、流程自动化、实施方法等）
第三天下午	14:00-17:00	客户团队针对其他三个流程，参考AWS提供的方法和最佳实践，讨论端到端的实施方法
	17:00-18:00	AWS结合客户团队的管理实践、讨论形成从易到难的实施路线图
第四天上午	09:00 - 11:00	在应用现代化架构的层面，理解AWS的方法论、最佳实践和参考架构
	11:00 – 13:00	客户应用架构、技术架构和运维人员讨论应用现代化得主要考量，AWS团队提供建议
第四天下午	14:00-15:00	BC Lab的快速总结
第五天上午	9:00-12:00	沙箱演练和系统演练
第五天下午	14：00-16:00	演练总结及高层汇报
	16:00-18:00	方法回复及下一步展望

CI/CD/CO方法，涉及文化、组织、流程、架构和工具的综合改进。为具象化改进，并帮助金融客户短期内看到效益，AWS重点推出Business Continuity Lab。BC lab分成三个阶段，双方各有四名专家，基于客户的突出问题开展工作。

客户 角色	职责
Director	高层指导
Owner	结果负责、资源协调
架构师	架构需求及设计
运维专家	流程需求及设计
开发专家	流程需求及设计

AWS 角色	职责
Director	高层指导
Owner	结果负责
首席顾问	流程质量和架构质量
技术专家	代码质量
项目经理	资源协调

Pre-Lab

1. 双方：
   - 确定Lab的目标和意义
   - 确定双方人员
   - 初步确定具体最影响业务连续性、最具备短期操作可能性的问题
2. 客户方：
   - 确定Lab时间、地点、形式
   - 提供必要的基础材料
3. AWS方：
   - 组织团队
   - 研读材料
   - 拟定方案（含验证code）

Field-Lab

为期五天（具体工作内容见下页）

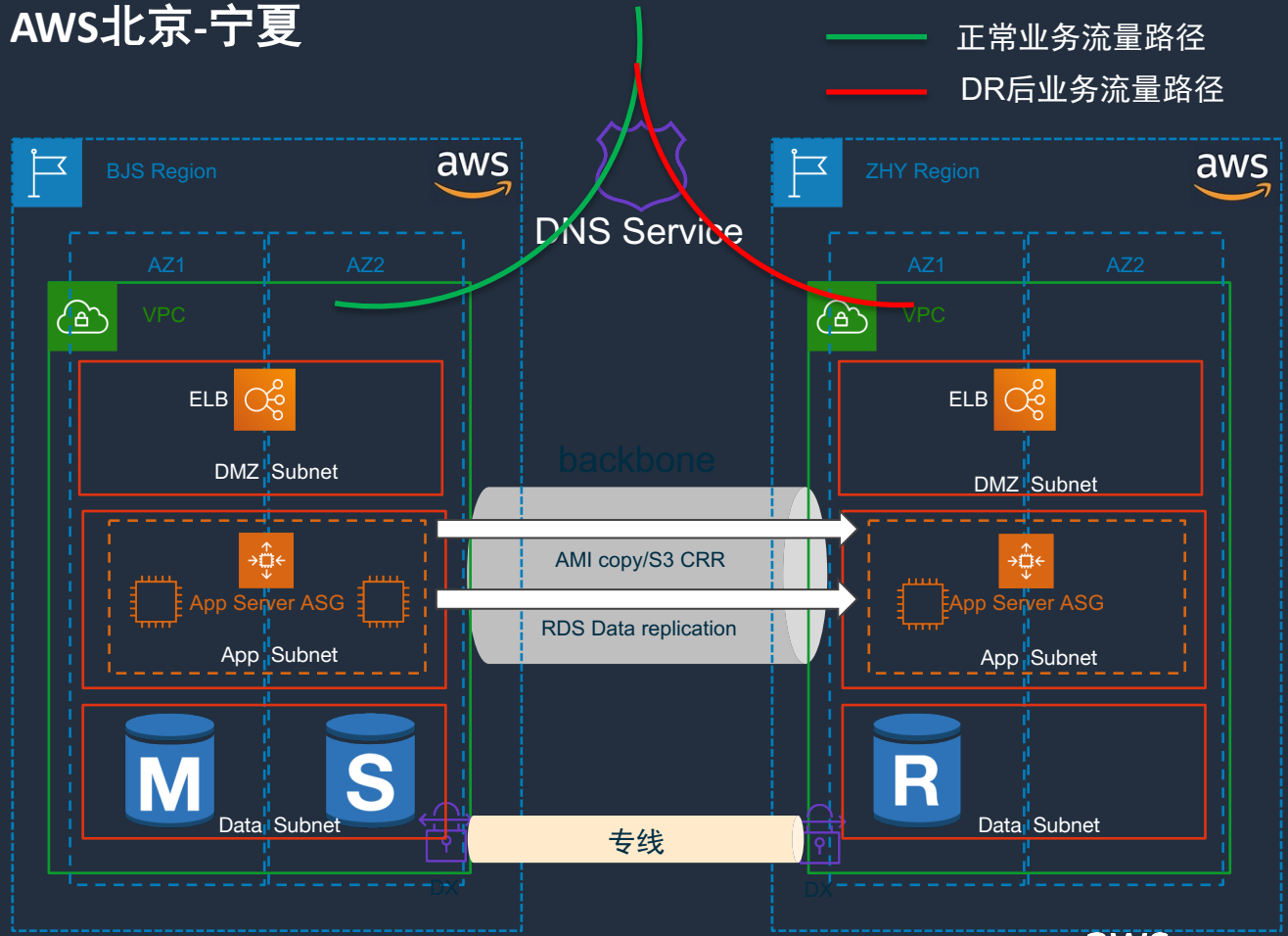
Post-Lab

1. 基于操作手册，完成真实场景的变更操作
2. 组织Roll-out流程的工作模式
3. 开展整体架构的调整

AWS多Region容灾架构：AWS北京-宁夏

具备完整功能的缩小版环境始终在灾备区域中运行。

在生产系统发生故障的情况下，备用环境将按比例放大以用于生产负载，并且将更改DNS记录以将所有流量路由到另一个区域



AWS 中国（宁夏）区域由西云数据运营
AWS 中国（北京）区域由光环新网运营