

ALIBABA CLOUD

阿里云

阿里云最佳实践
数据分析

文档版本：20220121

 阿里云

法律声明

阿里云提醒您在阅读或使用本文档之前仔细阅读、充分理解本法律声明各条款的内容。如果您阅读或使用本文档，您的阅读或使用行为将被视为对本声明全部内容的认可。

1. 您应当通过阿里云网站或阿里云提供的其他授权通道下载、获取本文档，且仅能用于自身的合法合规的业务活动。本文档的内容视为阿里云的保密信息，您应当严格遵守保密义务；未经阿里云事先书面同意，您不得向任何第三方披露本手册内容或提供给任何第三方使用。
2. 未经阿里云事先书面许可，任何单位、公司或个人不得擅自摘抄、翻译、复制本文档内容的部分或全部，不得以任何方式或途径进行传播和宣传。
3. 由于产品版本升级、调整或其他原因，本文档内容有可能变更。阿里云保留在没有任何通知或者提示下对本文档的内容进行修改的权利，并在阿里云授权通道中不时发布更新后的用户文档。您应当实时关注用户文档的版本变更并通过阿里云授权渠道下载、获取最新版的用户文档。
4. 本文档仅作为用户使用阿里云产品及服务的参考性指引，阿里云以产品及服务的“现状”、“有缺陷”和“当前功能”的状态提供本文档。阿里云在现有技术的基础上尽最大努力提供相应的介绍及操作指引，但阿里云在此明确声明对本文档内容的准确性、完整性、适用性、可靠性等不作任何明示或暗示的保证。任何单位、公司或个人因为下载、使用或信赖本文档而发生任何差错或经济损失的，阿里云不承担任何法律责任。在任何情况下，阿里云均不对任何间接性、后果性、惩戒性、偶然性、特殊性或刑罚性的损害，包括用户使用或信赖本文档而遭受的利润损失，承担责任（即使阿里云已被告知该等损失的可能性）。
5. 阿里云网站上所有内容，包括但不限于著作、产品、图片、档案、资讯、资料、网站架构、网站画面的安排、网页设计，均由阿里云和/或其关联公司依法拥有其知识产权，包括但不限于商标权、专利权、著作权、商业秘密等。非经阿里云和/或其关联公司书面同意，任何人不得擅自使用、修改、复制、公开传播、改变、散布、发行或公开发表阿里云网站、产品程序或内容。此外，未经阿里云事先书面同意，任何人不得为了任何营销、广告、促销或其他目的使用、公布或复制阿里云的名称（包括但不限于单独为或以组合形式包含“阿里云”、“Aliyun”、“万网”等阿里云和/或其关联公司品牌，上述品牌的附属标志及图案或任何类似公司名称、商号、商标、产品或服务名称、域名、图案标示、标志、标识或通过特定描述使第三方能够识别阿里云和/或其关联公司）。
6. 如若发现本文档存在任何错误，请与阿里云取得直接联系。

通用约定

格式	说明	样例
 危险	该类警示信息将导致系统重大变更甚至故障，或者导致人身伤害等结果。	 危险 重置操作将丢失用户配置数据。
 警告	该类警示信息可能会导致系统重大变更甚至故障，或者导致人身伤害等结果。	 警告 重启操作将导致业务中断，恢复业务时间约十分钟。
 注意	用于警示信息、补充说明等，是用户必须了解的内容。	 注意 权重设置为0，该服务器不会再接受新请求。
 说明	用于补充说明、最佳实践、窍门等，不是用户必须了解的内容。	 说明 您也可以通过按Ctrl+A选中全部文件。
>	多级菜单递进。	单击设置> 网络> 设置网络类型。
粗体	表示按键、菜单、页面名称等UI元素。	在结果确认页面，单击确定。
Courier字体	命令或代码。	执行 <code>cd /d C:/window</code> 命令，进入Windows系统文件夹。
斜体	表示参数、变量。	<code>bae log list --instanceid</code> <i>Instance_ID</i>
[] 或者 [a b]	表示可选项，至多选择一个。	<code>ipconfig [-all -t]</code>
{ } 或者 {a b}	表示必选项，至多选择一个。	<code>switch {active stand}</code>

目录

1.游戏数据运营融合分析	06
2.基于flowlog的流量分析	08
3.自建Hive数据仓库跨版本迁移到阿里云Databricks数据洞察	09
4.基于Elasticsearch的订单检索加速最佳实践	10
5.互联网、游戏等行业实时数仓构建	11
6.互联网、电商Elasticsearch搜索	12
7.互联网电商行业离线大数据分析	13
8.互联网电商与游戏行业实时BI分析	14
9.自建Hadoop迁移MaxCompute	15
10.自建Hadoop迁移到阿里云EMR	16
11.自建Hive数仓迁移到阿里云EMR	17
12.云上数据OLAP+OLTP	18
13.基于Flink+ClickHouse构建实时游戏数据分析	19
14.大数据近实时数据投递MaxCompute	20
15.基于湖仓一体架构使用MaxCompute对OSS湖数据分析预测最佳实践	21
16.电商网站智能推荐	22
17.电商网站数据埋点及分析	24
18.微服务架构日志采集运维管理	25
19.抢占式ECS搭建离线大数据分析集群	27
20.SLS多云日志采集、处理及分析	28
21.EMR弹性低成本离线大数据分析	29
22.互联网、电商行业实时大数据分析	30
23.互联网、电商行业离线大数据分析	31
24.基于MaxCompute的大数据BI分析	32
25.EMR本地盘实例大规模数据集测试	34
26.EMR弹性低成本离线大数据分析	35

27.互联网、电商及游戏行业实时BI分析	36
28.低成本基于抢占式ECS实例进行离线大数据分析	37
29.DTS数据同步集成MaxCompute数仓	39
30.Spark on ECI大数据分析	40
31.数据湖：在线学习场景数据分析	42
32.互联网、游戏等行业实时数仓构建	43
33.基于Flink的咨询场景实时数仓	44
34.基于Dataworks的大数据一站式开发及数据治理	45
35.利用交互分析（Hologres）进行数据查询解决方案	47
36.MaxCompute湖仓一体方案	49

1. 游戏数据运营融合分析

针对游戏行业数据分析实时性高、结构化和非结构化数据融合需求，构建游戏数据运营融合分析一体化架构。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践](#)

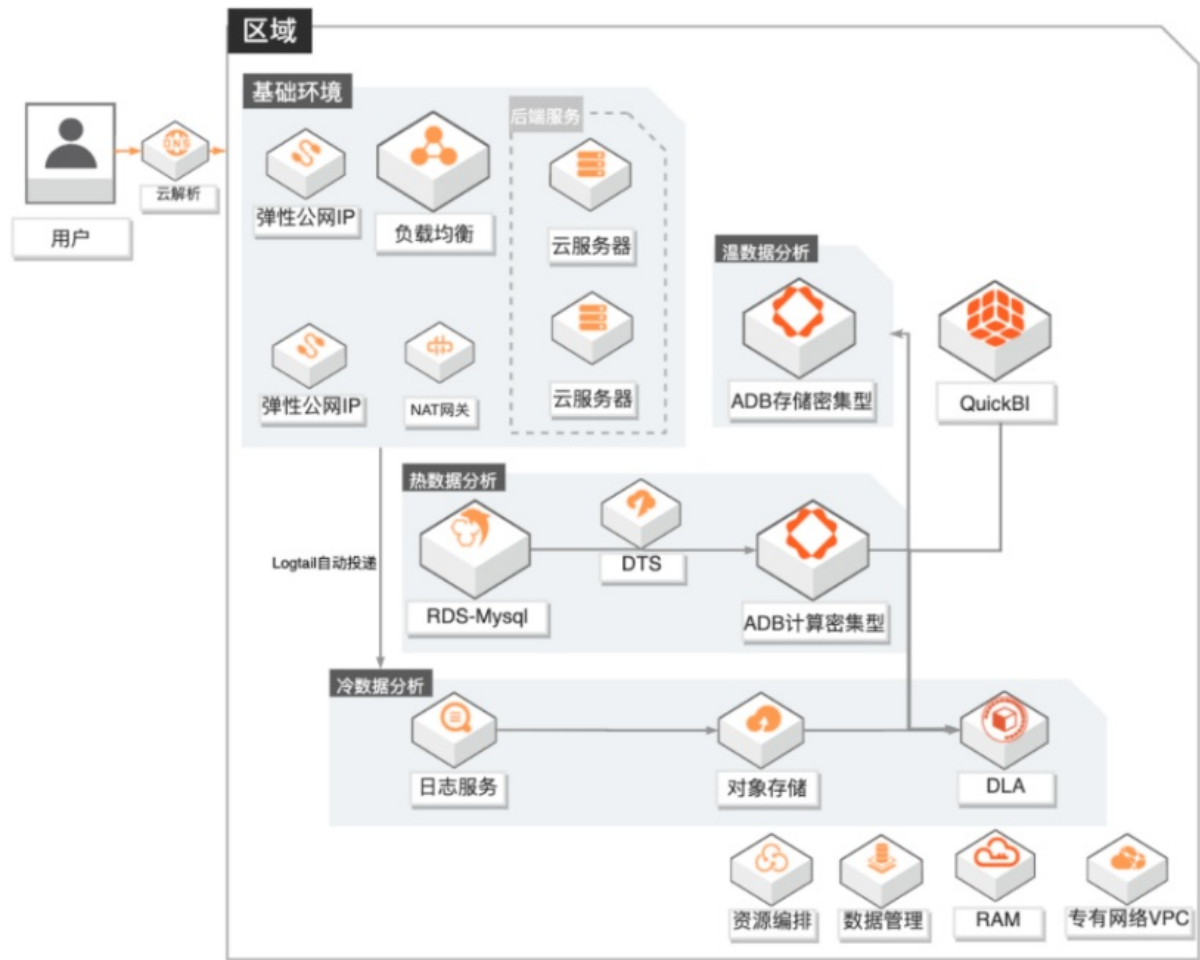
场景描述

- 游戏行业有结构化和非结构化数据融合分析需求的客户。
- 游戏行业有数据实时分析需求的客户，无法接受T+1延迟。
- 对数据成本有一定诉求的客户，希望物尽其用尽量优化成本。
- 其他行业有类似需求的客户。

解决问题

- 秒级实时分析：依托ADB计算密集型实例，秒级监控DAU等数据，为广告投放效果提供有力的在线决策支撑。
- 高效数据融合分析：打通结构化和非结构化数据，支撑产品体验分析；广告买量投放效果实时（分钟级）分析，渠道的评估更准确。
- 低成本：DLA融合冷数据分析+ADB存储密集型温数据分析+ADB计算密集型热数据分析，在满足各种分析场景需求的同时，有效地降低的客户的总体使用成本。
- 学习成本低：DLA和ADB兼容标准SQL语法，无需额外学习其他技术。

部署架构



2.基于flowlog的流量分析

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多最佳实践](#)

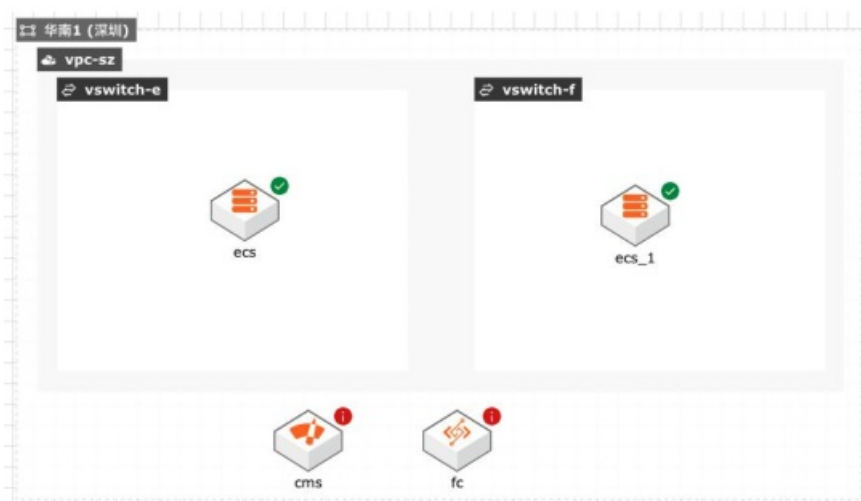
场景描述

用来分析流量异常的场景，如VPC内ECS异常流量查看、NAT网关高流量ECS排查、CEN东西向流量查看、混合云专线流量查看等场景，排查问题原因，检查网络配置规则，提高运维管理效率。

解决问题

- 查看相同VPC下不同ECS的互访流量并排查问题
- 排查公网流量占比较高的ECS问题
- 排查不同VPC间流量互访问题
- 自动化开启和关闭流日志，提高运维效率，降低使用成本。

业务架构



3.自建Hive数据仓库跨版本迁移到阿里云Databricks数据洞察

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多最佳实践](#)

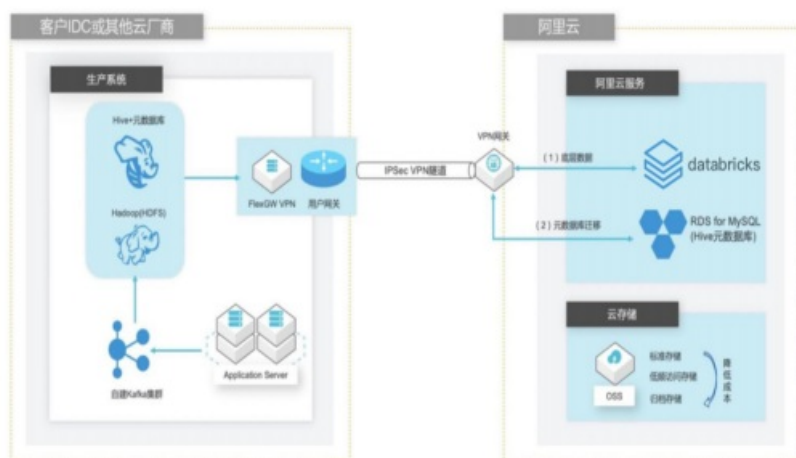
场景描述

客户在IDC或者公有云环境自建Hadoop集群构建数据仓库和分析系统，购买阿里云Databricks数据洞察集群之后，涉及到数仓数据和元数据的迁移以及Hive版本的订正更新。

解决问题

- Hive数仓数据迁移OSS方案
- Hive元数据库迁移阿里云RDS方案
- Hive跨版本迁移到Databricks数据洞察使用Delta表查询以提高查询效率

方案架构



4.基于Elasticsearch的订单检索加速最佳实践

本实践基于阿里云Elasticsearch作为二级索引库，数据集成产品提供Binlog实时订阅，实时解析、增量数据实时更新及二级索引库之间的数据实时同步，为数据库提供“能力增益”，不仅能从根本解决主库抗压问题，提升稳定性；同时支持高效率、高性能、高弹性、低成本、多复杂场景的检索加速服务。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多最佳实践](#)

场景描述

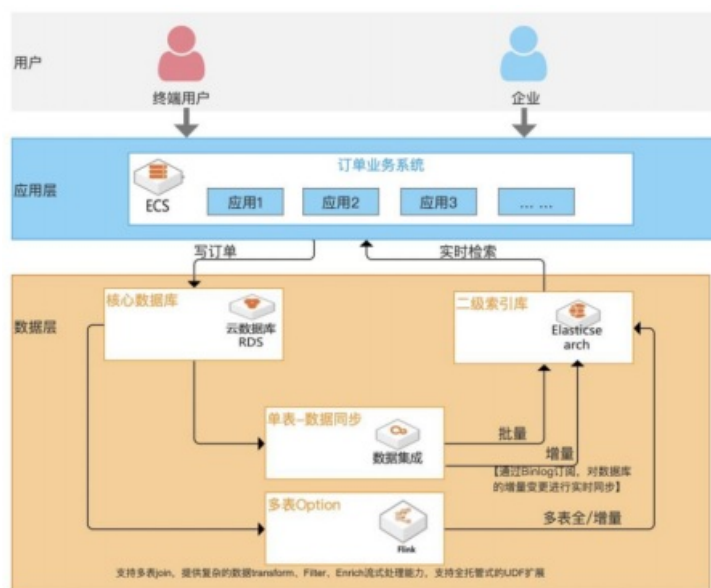
随着企业信息化程度越来越高，核心业务数据存储在传统关系型数据库中不可避免地会遇到一个问题：单表记录不断增多，数据检索速度会变慢，尤其是对中文的模糊查询（建立普通索引完全不起作用）。虽然数据库自身在不断完善，但效果有限且没办法灵活扩展，复杂场景无法应对。

本实践基于阿里云Elasticsearch作为二级索引库，数据集成产品提供Binlog实时订阅，实时解析、增量数据实时更新及二级索引库之间的数据实时同步，为数据库提供“能力增益”，不仅能从根本解决主库抗压问题，提升稳定性；同时支持高效率、高性能、高弹性、低成本、多复杂场景的检索加速服务

解决问题

- 单表容量瓶颈查询压力大，主库稳定性风险高。
- 数据库查询并发性弱
- 数据库横向扩展能力弱
- 复杂查询无法满足（全文、模糊查询、聚合查询、多维过滤等复杂场景）。

方案架构



直达最佳实践

更多最佳实践

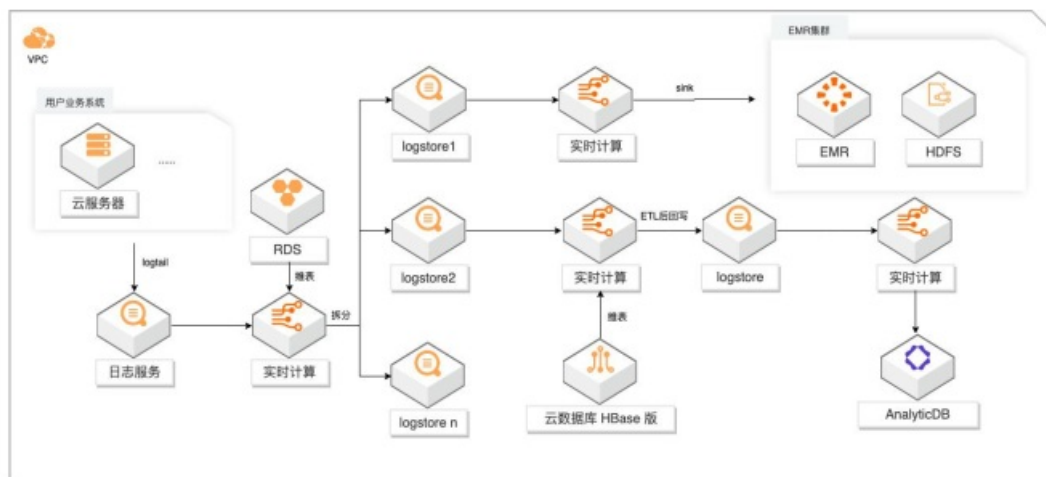
场景描述

互联网、游戏等行业客户基于Flink构建实时数仓，通过Logtail采集日志，使用Flink实现对日志的拆分、处理，根据业务需求写入不同的数据存储，实时分析使用ADB存储、离线分析使用HDFS存储，构建高效、稳健的实时数据服务。

解决问题

- 日志采集、处理及分析
- 日志使用Flink拆分到不同logstore
- 日志使用Flink实时写入HDFS
- 日志数据实时ETL
- 日志实时写入ADB

方案架构



6.互联网、电商Elasticsearch搜索

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多最佳实践](#)

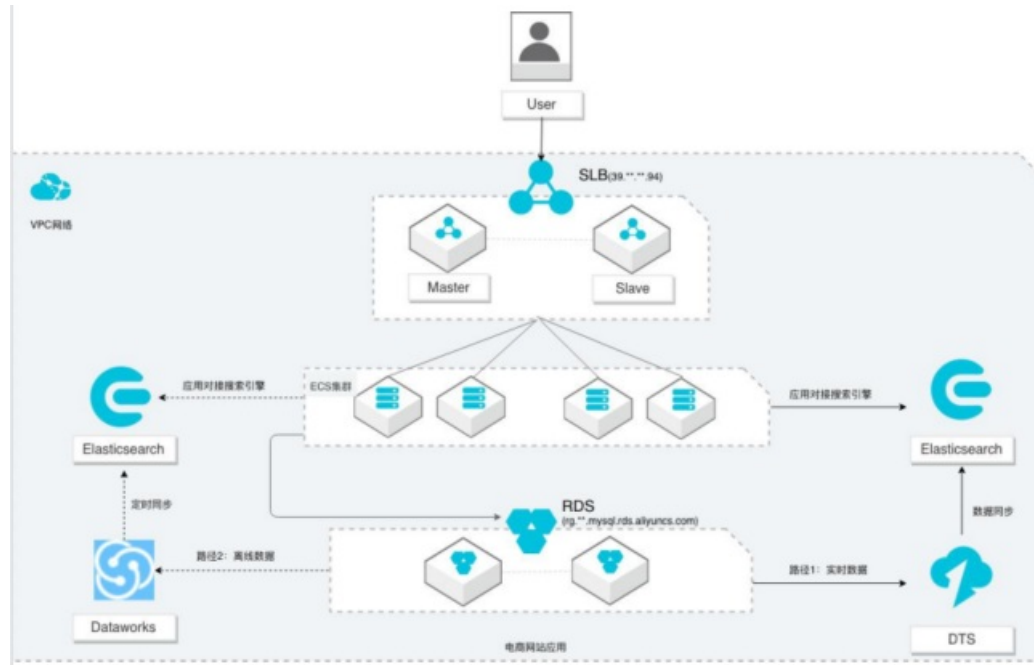
场景描述

每一个生活在互联网中的用户，每天都在经历各种各样的“搜索”，查找电商网站商品、信用卡账单、查电子发票、查附近的餐厅酒店、查偶像、查交通等等。相对于传统的关系型数据库，Elasticsearch只需要几毫秒的时间，即可查询PB级数据并从中找到匹配信息。利用Elasticsearch高可用性和易用性，能够快速处理网站、APP丢给它的文本、数字、日期、IP以及地理数据。

解决问题

- 电商网站商品、订单、物流等搜索
- 保险行业保单查询
- 互联网教育行业课程搜索
- O2O行业地理位置搜索等

方案架构



7.互联网电商行业离线大数据分析

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多最佳实践](#)

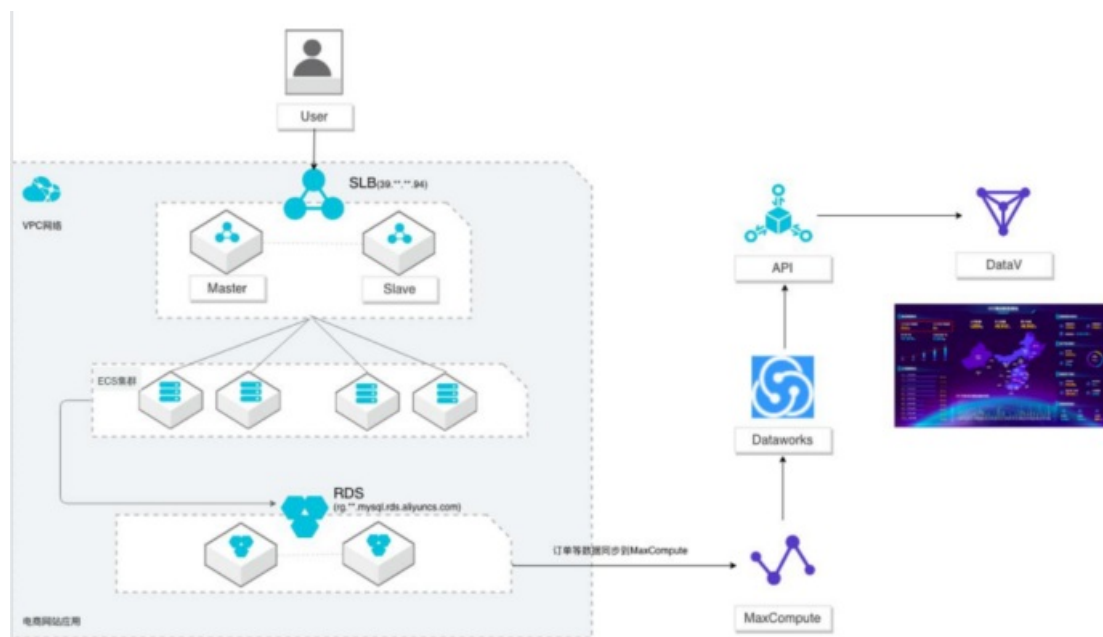
场景描述

电商网站销售数据通过大数据分析后将业务指标数据在大屏幕上展示，如销售指标、客户指标、销售排名、订单地区分布等。大屏上销售数据可视化动态展示，效果震撼，触控大屏支持用户自助查询数据，极大地增强数据的可读性。

解决问题

- 电商网站数据看板
- 全国/全球业务态势分析
- 互联网金融风险数据监控

方案架构



8. 互联网电商与游戏行业实时BI分析

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多最佳实践](#)

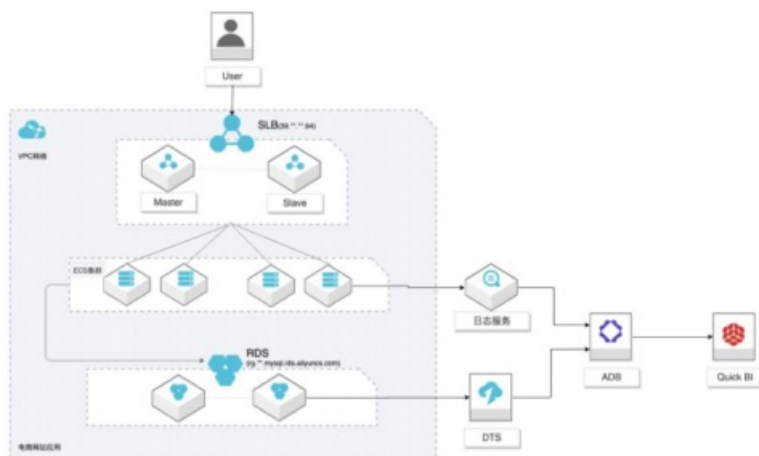
场景描述

本文以电商行业为例，将业务数据和日志数据同步到ADB，之后通过QuickBI做实时可视化分析。相对于传统的关系型数据库，阿里云分析型数据库MySQL版只需要几毫秒的时间，即可查询PB级数据并从中找到匹配信息。

解决问题

- 互联网行业、电商、游戏行业等网站、App、小程序应用内BI分析场景。
- 在线运营、运营指标实时化分析等场景。
- 可扩展到各类网站BI分析场景使用。

方案架构



9.自建Hadoop迁移MaxCompute

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多最佳实践](#)

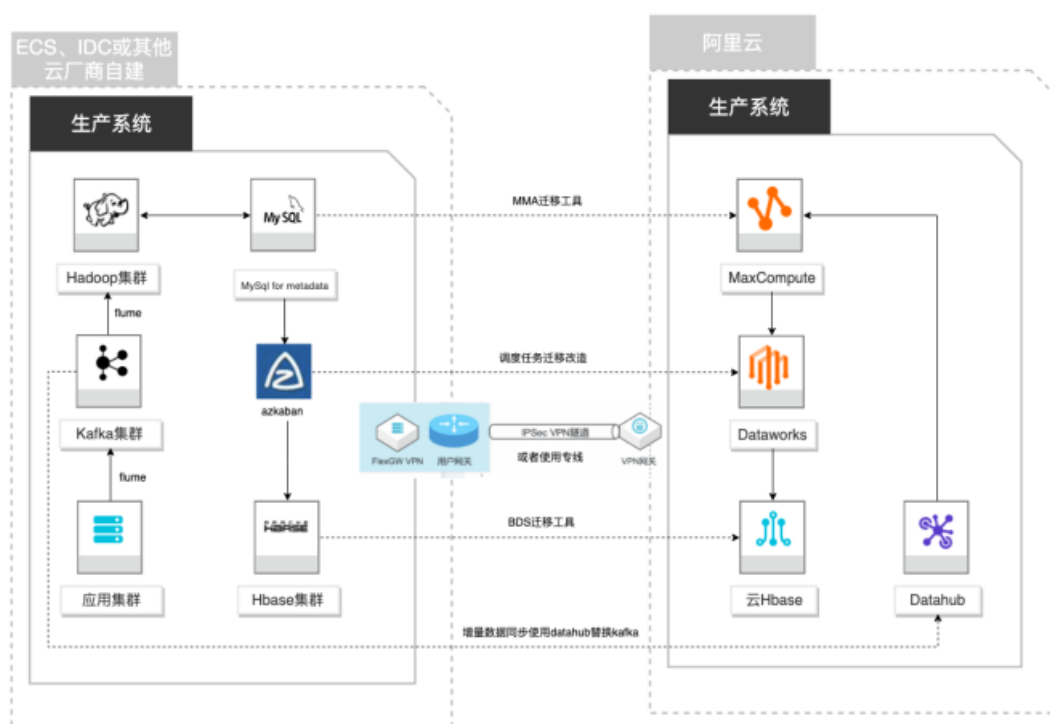
场景描述

客户基于ECS、IDC自建或在友商云平台自建了大数据集群，为了降低企业大数据计算平台的成本，提高大数据应用开发效率，更有效保障数据安全，把大数据集群的数据、作业、调度任务以及业务数据库整体迁移到MaxCompute和其他云产品。

解决问题

- 自建Hadoop集群搬迁到MaxCompute
- 自建Hbase集群搬迁到云Hbase
- 自建Kafka或应用数据准实时同步到MaxCompute
- 自建Azkaban任务迁移到Dataworks任务

方案架构



10. 自建Hadoop迁移到阿里云EMR

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多最佳实践](#)

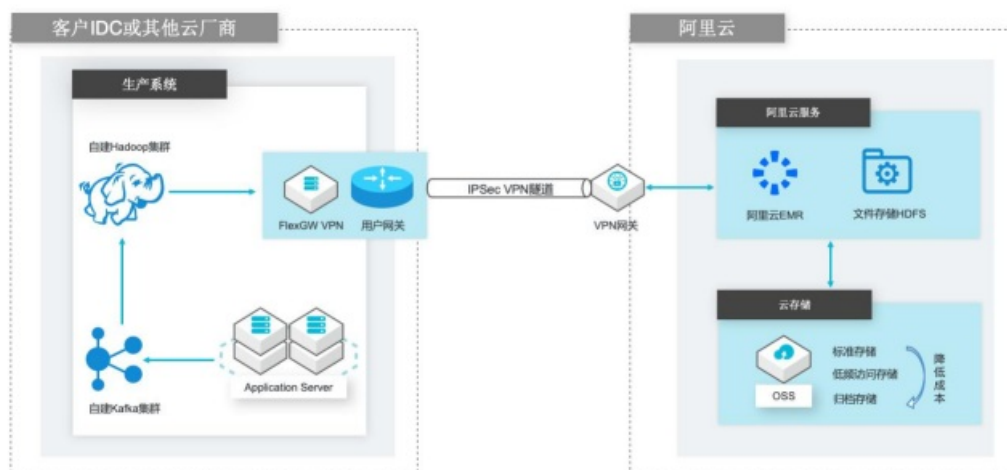
场景描述

- 场景1：自建Hadoop集群数据（HDFS）迁移到阿里云EMR集群的HDFS文件系统；
- 场景2：自建Hadoop集群数据（HDFS）迁移到计算存储分离架构的阿里云EMR集群，以OSS和jindoFS作为EMR集群的后端存储。

解决问题

- 客户自建Hadoop迁移到阿里云EMR集群的技术方案
- 基于IPSecVPN隧道构建安全和低成本数据传输链路

方案架构



11.自建Hive数仓迁移到阿里云EMR

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多最佳实践](#)

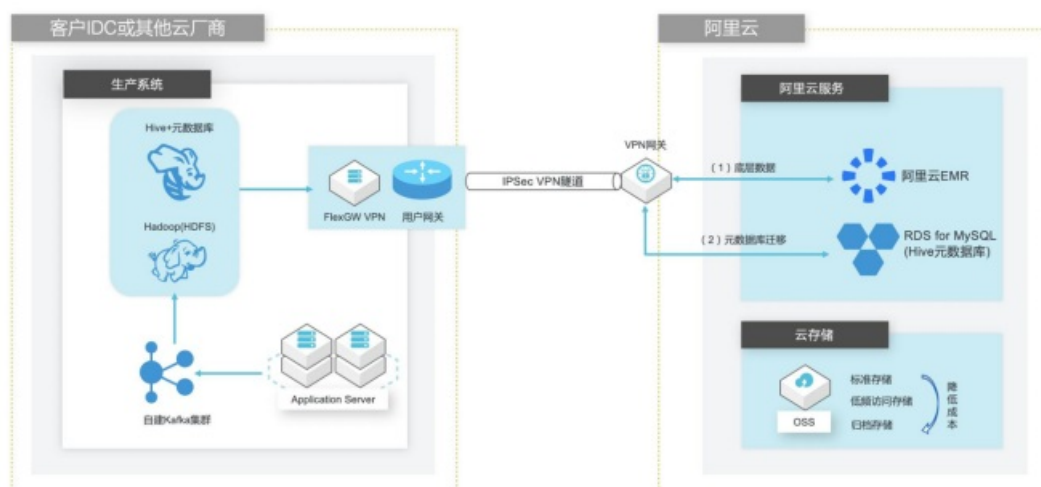
场景描述

客户在IDC或者公有云环境自建Hadoop集群构建数据仓库和分析系统，购买阿里云EMR集群之后，涉及到将数据仓库和Hive元数据的数据库迁移上云。目前主流Hive数据仓库迁移场景为1.x版本迁移到阿里云EMR(Hive2.x版本)，涉及到数据订正更新步骤。

解决问题

- Hive数据仓库的数据迁移方案
- Hive元数据库的迁移方案
- Hive跨版本迁移后的数据订正
- 的管控权

方案架构



12.云上数据OLAP+OLTP

本最佳实践提供数据库上云后OLTP+OLAP分离架构的最佳实践，并为客户提供完整的OLAP、OLTP数据库系统压测对比方式。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践。](#)

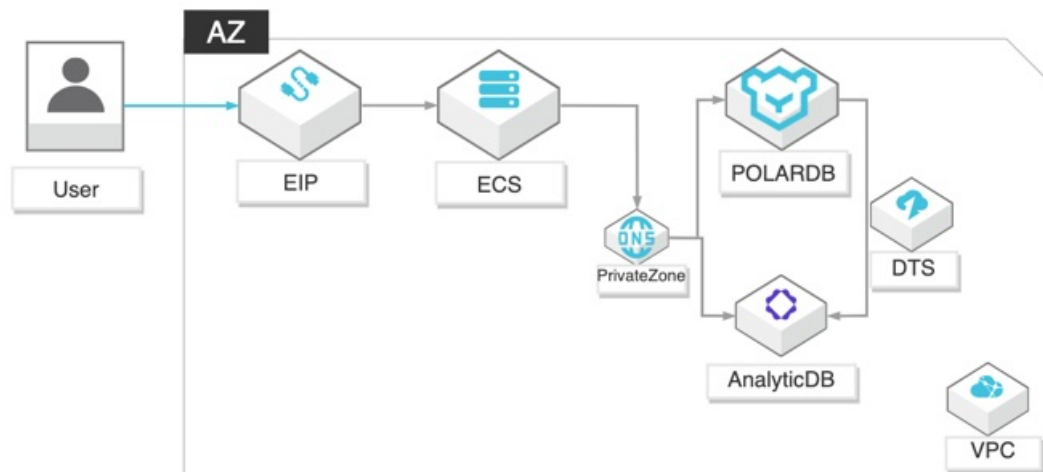
场景描述

- 混合型HTAP数据库需求：如ORACLE数据库改造上云，云上数据库方案选型。
- OLTP系统查询慢，存在分析型场景和瓶颈的客户。

解决的问题

- 实时性强：利用DTS高实时将业务库数据变更同步到ADB中，供分析查询业务查询，与线上业务数据延迟低。
- 提供阿里云强大的HTAP能力：OLTP和OLAP需求通过PolarDB和ADB分治，解决客户HTAP的复杂需求。
- 提供完备的OLAP、OLTP压测的工具和方法，指导客户压测对比，便于客户POC测试和数据库选型。

部署架构图



13.基于Flink+ClickHouse构建实时游戏数据分析

本实践将模拟某款游戏线上的用户行为数据，基于日志文件和实时产生的行为数据，我们可以做后续的实时用户行为分析实验，体验这套解决方案的落地实践。

直达最佳实践

[点击查看最佳实践详情](#)

直达最佳实践

[点击查看最佳实践详情](#)

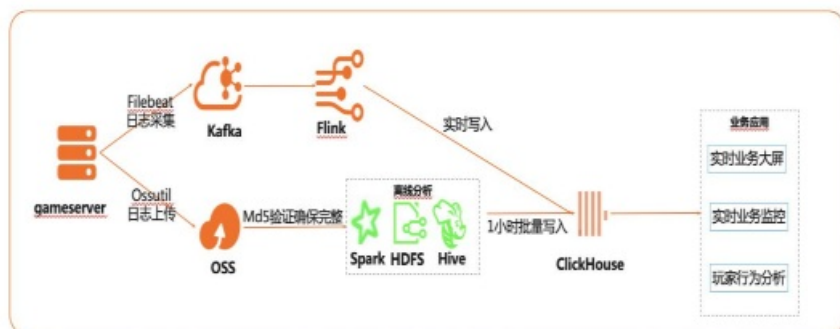
场景描述

本实践介绍如何快速收集海量用户行为数据，实现秒级响应的实时用户行为分析，并通过实时流计算、云数据库ClickHouse等技术进行深度挖掘和分析，得到用户特征和画像，实现个性化系统推荐服务。

方案优势

- 通过云数据库ClickHouse替换原有Presto数仓，对比开源Presto性能提升20倍。
- 利用云数据库ClickHouse极致分析性能，千亿级数据分析从10分钟缩短到30秒。
- 云数据库ClickHouse批量写入效率高，支持业务高峰每小时230亿的用户数据写入。
- 云数据库ClickHouse开箱即用，免运维，全球多Region部署，快速支持新游戏开服。

方案架构



14.大数据近实时数据投递MaxCompute

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多最佳实践](#)

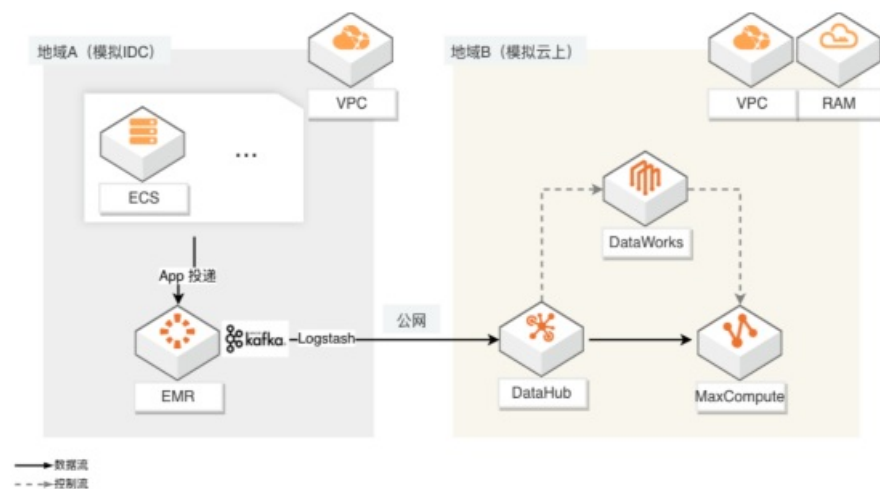
场景描述

本文介绍离线大数据场景使用MaxCompute构建云上近实时数仓，打通云下数据上云链路，解决数据复杂类型支持和动态分区问题，满足高级数据处理需求的最佳实践。

解决问题

- 混合云环境下，现有业务系统零改造，打通数据上云链路。
- 使用UDF实现复杂数据类型转换和数据动态分区。
- 使用DataWorks配置周期调度业务流程，数据自动入仓。
- 借助MaxCompute优化计算引擎，实现降本增效。

方案架构



15.基于湖仓一体架构使用MaxCompute对OSS湖数据分析预测最佳实践

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多最佳实践](#)

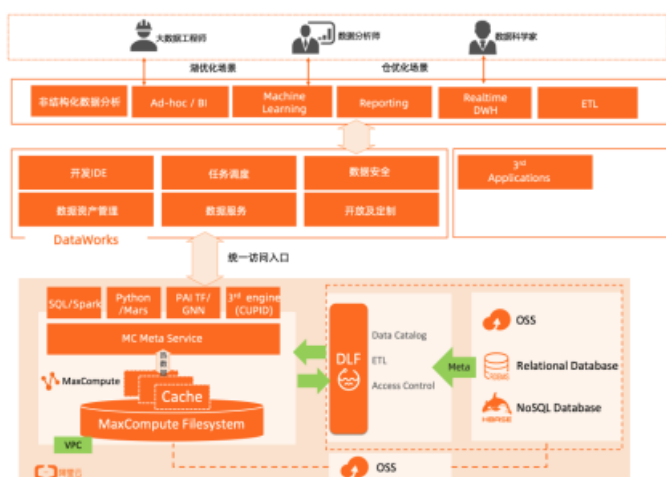
场景描述

数据湖和数据仓库是当前大数据技术条件下构建分布式系统的两种数据架构设计取向，数据湖偏向灵活性，数据仓库侧重成本、性能、安全、治理等企业级特性。但是数据湖和数据仓库的边界正在慢慢模糊，数据湖自身的治理能力、数据仓库延伸到外部存储的能力都在加强。在这样的背景之下，阿里云MaxCompute率先提出湖仓一体，为业界和用户展现了一种数据湖和数据仓库互相补充，协同工作的架构。这样的架构同时为用户提供了数据湖的灵活性和数据仓库的诸多企业级特性，将用户使用大数据的总体拥有成本进一步降低。

方案优势

- 融合数据湖和数据仓库优势，在灵活性和效率上找到最佳平衡。
- MaxCompute在SQL上做了大量优化与能力沉淀，可提高SQL运行性能，降低计算成本。基于集群学习PAI封装出多种贴近业务场景的算法服务，满足更多的业务需求。
- MaxCompute云原生的弹性资源和EMR集群资源形成互补，两套体系之间进行资源的削峰填谷，不仅减少作业排队，且降低整体成本。

方案架构



16. 电商网站智能推荐

电商网站向用户推荐带有商品属性物品以促进交易，提升购买率和转化率。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践](#)

场景描述

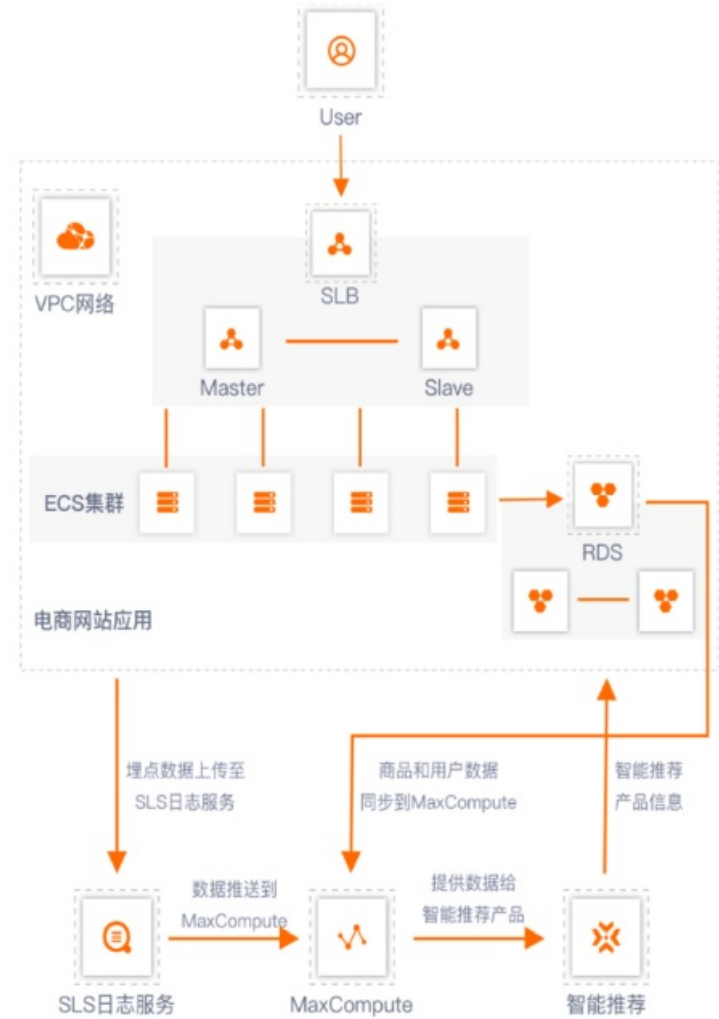
基于阿里巴巴领先的大数据和人工智能技术，结合在电商行业的多年积累，为开发者提供个性化推荐服务，提升商品购买率和转化率。

解决问题

- 猜你喜欢/个性化内容瀑布流展示
- 在商品页提供产品的相关推荐
- 网站首页提供焦点图推荐和热门推荐
- 可扩展到新闻资讯、视频直播和社交领域

部署架构图

本实践以电商网站为例，使用日志服务采集日志，RDS作为后端数据存储服务，MaxCompute作为数据仓库并使用DataWorks进行数据同步和处理，使用智能推荐产品搭建电商网站智能推荐系统。在电商行业，需要向用户推荐的物品带有商品属性（物流信息，售卖信息），可以引导直接交易，提升购买率和转化率，适合首页推荐、信息流等相关场景。



17. 电商网站数据埋点及分析

对网站用户行为的每一个事件对应的位置进行埋点通过SDK上报，汇总数据进行分析以推动产品优化及指导运营。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践。](#)

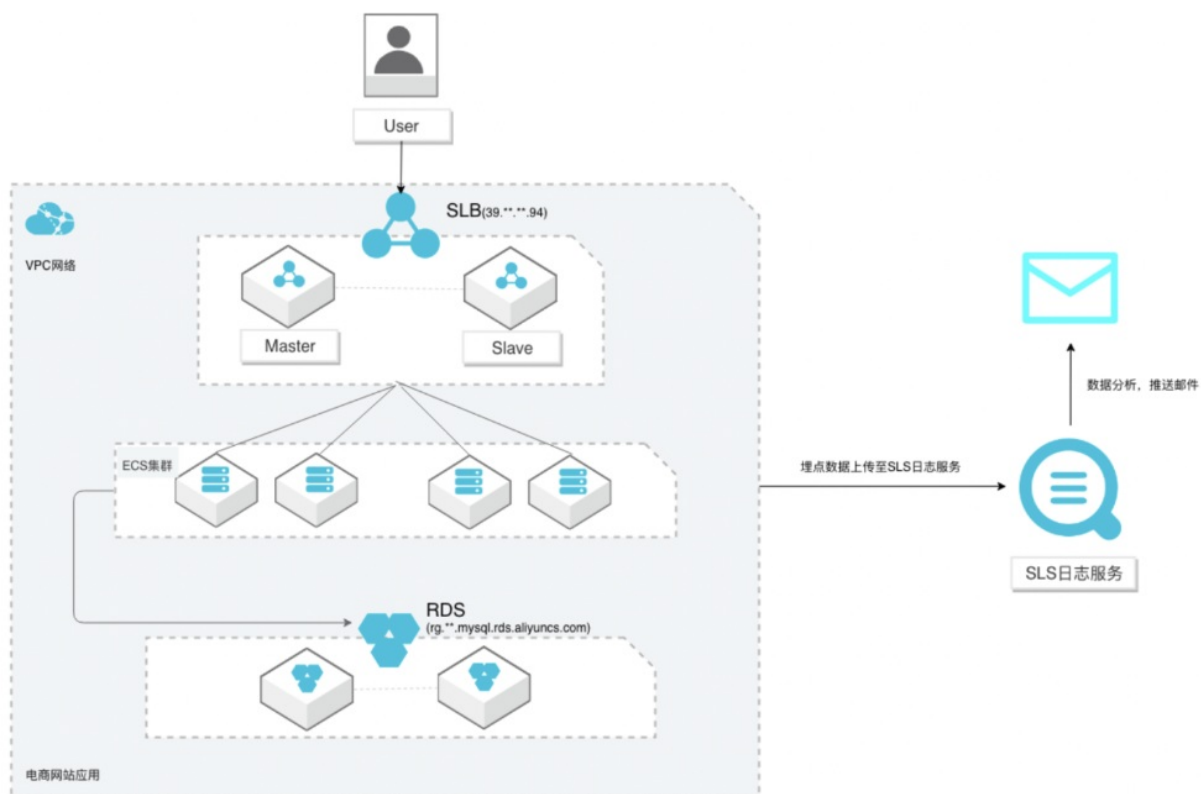
场景描述

数据埋点是数据产品经理、数据运营以及数据分析师，基于业务需求（例如：CPC点击付费广告中，统计每一个广告位的点击次数）和产品需求（例如：推荐系统中，推荐商品的曝光次数以及点击的人数）对用户行为的每一个事件对应的位置进行开发埋点，并通过SDK上报埋点的数据结果，记录数据汇总后进行分析，推动产品优化或指导运营。

解决问题

- 电商网站广告位效果统计分析
- 电商网站推荐商品曝光、点击、购买等行为统计分析
- 电商网站用户分布分析
- 电商网站页面热点图分析等

部署架构图



18.微服务架构日志采集运维管理

微服务应用化过程的日志采集运维管理，解决用户微服务改造过程中日志采集处理分析痛点。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践。](#)

场景描述

阿里云日志服务（SLS）结合Kubernetes容器集群日志特点以及应用场景，提供全方位的容器微服务应用环境下的日志采集、处理以及分析的实践解决方案。

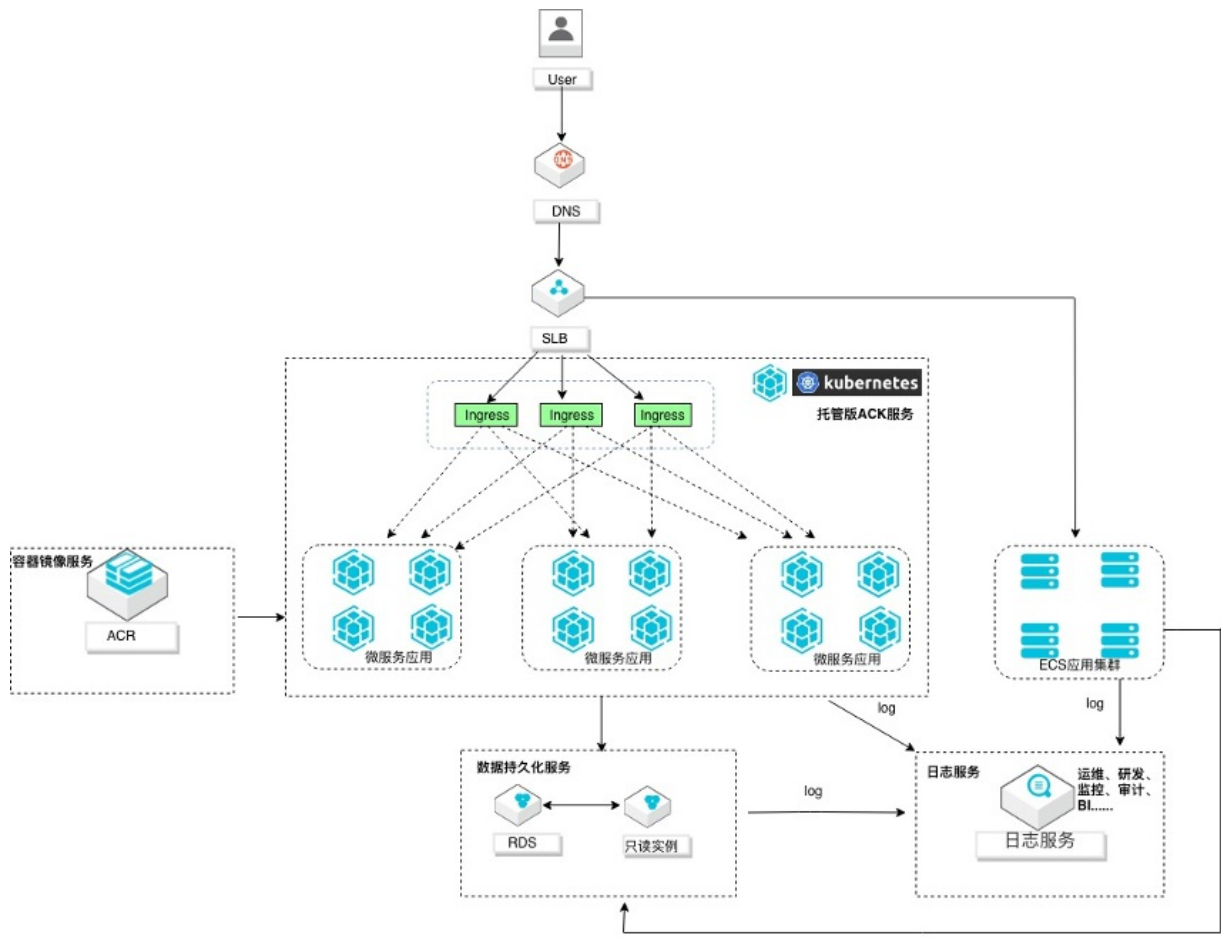
方案优势

- 日志采集配置灵活，支持多种日志形态，日志采集与Kubernetes完美集成
- 日志采集Agent占用资源少
- 安装部署方便，运维成本低
- 便捷实用，支持多种日志智能分析能力

解决问题

- 容器微服务架构下日志采集困难
- 日志采集Agent资源消耗多
- 运维日志平台系统成本代价高分布分析
- 日志分析方式少，日志分析不方便

部署架构图



19. 抢占式ECS搭建离线大数据分析集群

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多最佳实践](#)

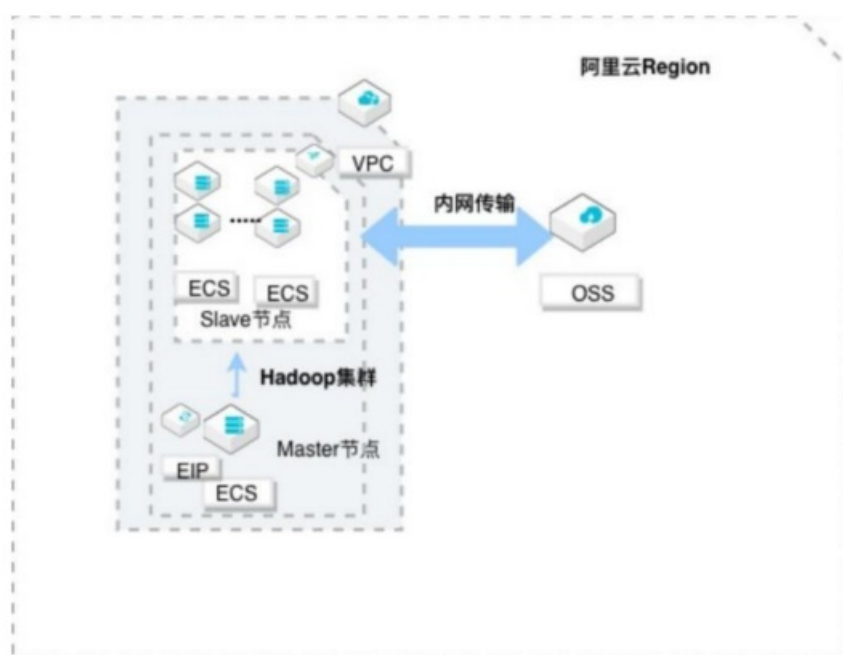
场景描述

基于阿里云的抢占式ECS实例以及弹性伸缩能力，同时结合阿里云对象存储OSS服务搭建离线大数据分析集群，实现计算与存储分离架构，同时降低计算资源成本和提升计算资源利用效率同时拥有海量存储能力。

解决问题

- 离线大数据分析集群计算与存储资源成本管控困难
- 计算资源弹性能力不足
- 大数据量大导致容量不足
- 大数据分析集群云资源性能优化

方案架构



20.SLS多云日志采集、处理及分析

多云平台服务器上采集日志写入到阿里云日志服务，帮助提升运维、运营效率。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践。](#)

场景描述

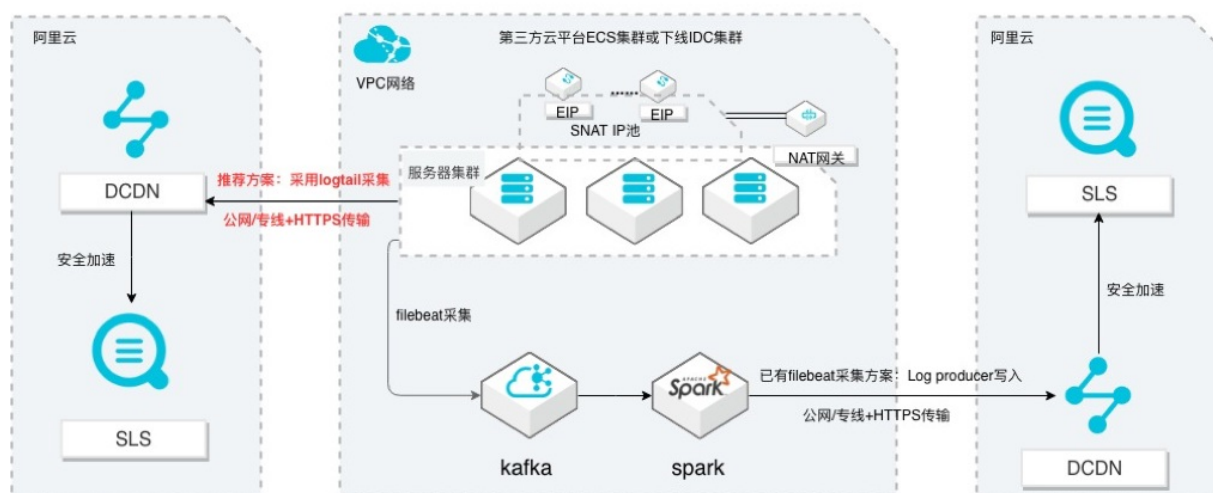
从第三方云平台或线下IDC服务器上采集日志写入到阿里云日志服务，通过日志服务进行数据分析，帮助提升运维、运营效率，建立DT时代海量日志处理能力。

针对未使用其他日志采集服务的用户，推荐在其他云或线下服务器安装logtail采集并使用HTTPS安全传输；针对已使用其他日志采集工具并且已有日志服务需要继续服务的情况，可以通过Log producer SDK写入日志服务。

解决问题

- 第三方云平台或线下IDC客户需要使用阿里云日志服务生态的用户
- 第三方云平台或线下IDC服务器已有完整日志采集、处理及分析的用户

部署架构图



21.EMR弹性低成本离线大数据分析

利用阿里云日志服务、存储服务以及EMR构建弹性低成本大数据日志离线分析系统。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践](#)

场景描述

基于阿里云的E-MapReduce(EMR)、对象存储OSS、日志服务SLS、抢占式ECS实例，构建弹性、低成本的计算与存储分离架构的海量离线大数据日志分析系统。

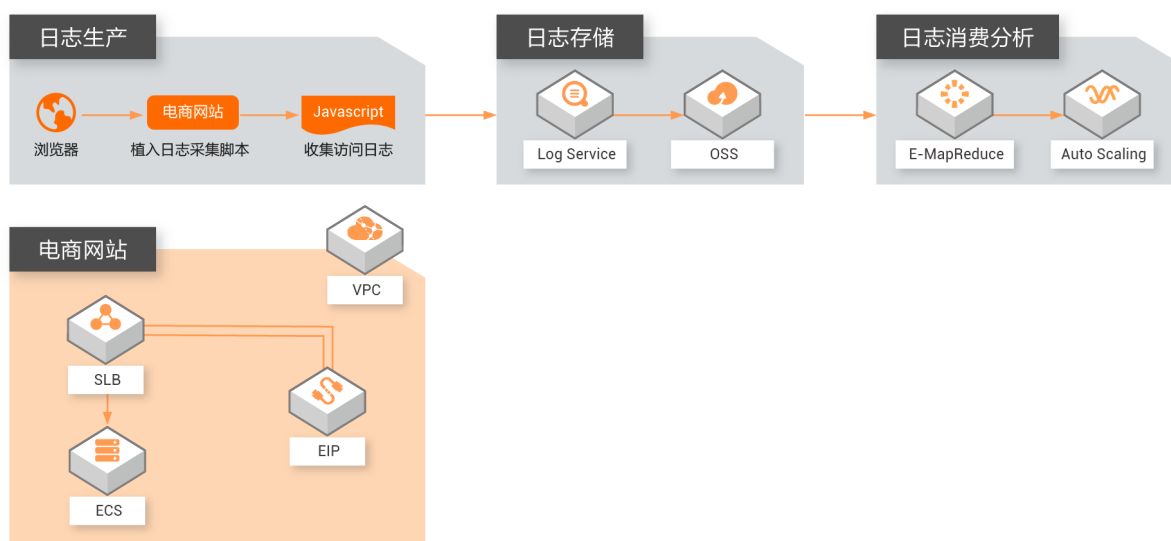
方案优势

- 弹性低成本、集群快速部署
- 支持多种大数据计算模式
- 无缝对接开源生态
- 一站式管理平台

解决问题

- 大数据平台运维管理成本高
- 计算资源弹性能力不足
- 海量日志数据下的存储容量不足
- 离线分析场景下的计算资源成本控制

部署架构图



22.互联网、电商行业实时大数据分析

使用阿里云服务实现电商网站购物数据实时分析后在大屏幕上展示，极大地增强数据的可读性。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践](#)

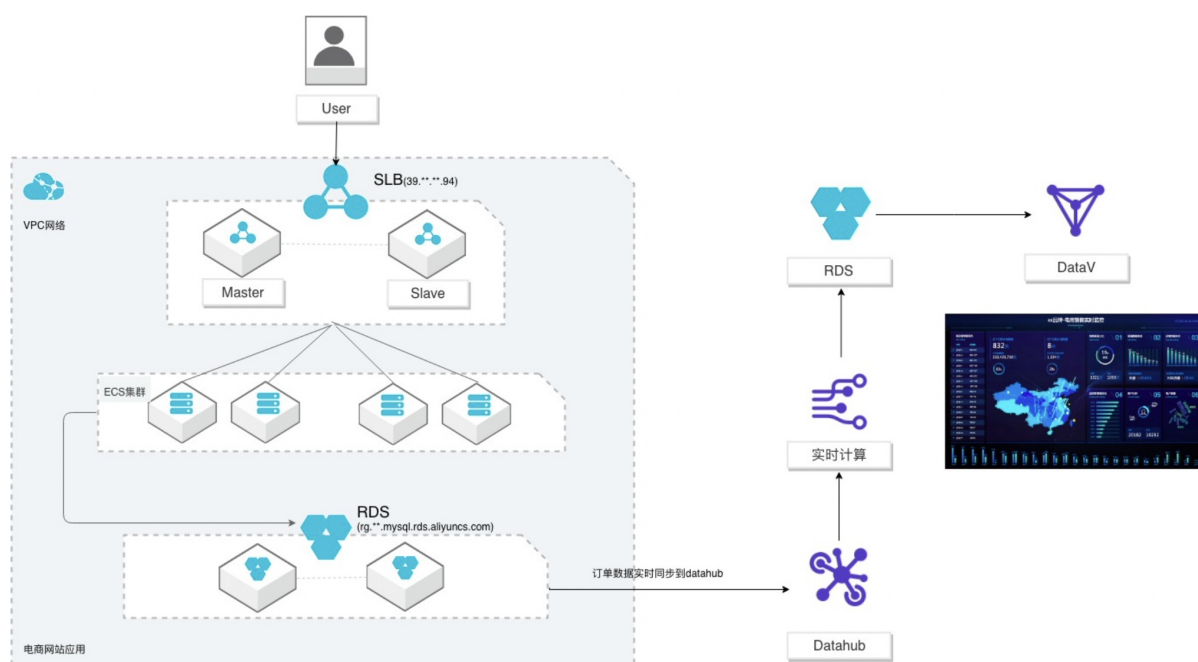
场景描述

电商网站购物数据通过实时大数据分析后在大屏幕上展示，如购买人群热销商品排名、购买人群区域分布、购买人群年龄分布等。大屏上全国上亿买家的订单数据可视化动态展示，效果震撼；触控大屏支持用户自助查询数据，极大增强数据的可读性。

解决问题

- 电商、营销大屏实时数据展示
- 内容、直播大屏实时数据展示
- 教育、金融大屏实时数据展示

部署架构图



23.互联网、电商行业离线大数据分析

使用阿里云服务实现电商网站销售数据离线分析后在大屏幕上展示，极大地增强数据的可读性。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践](#)

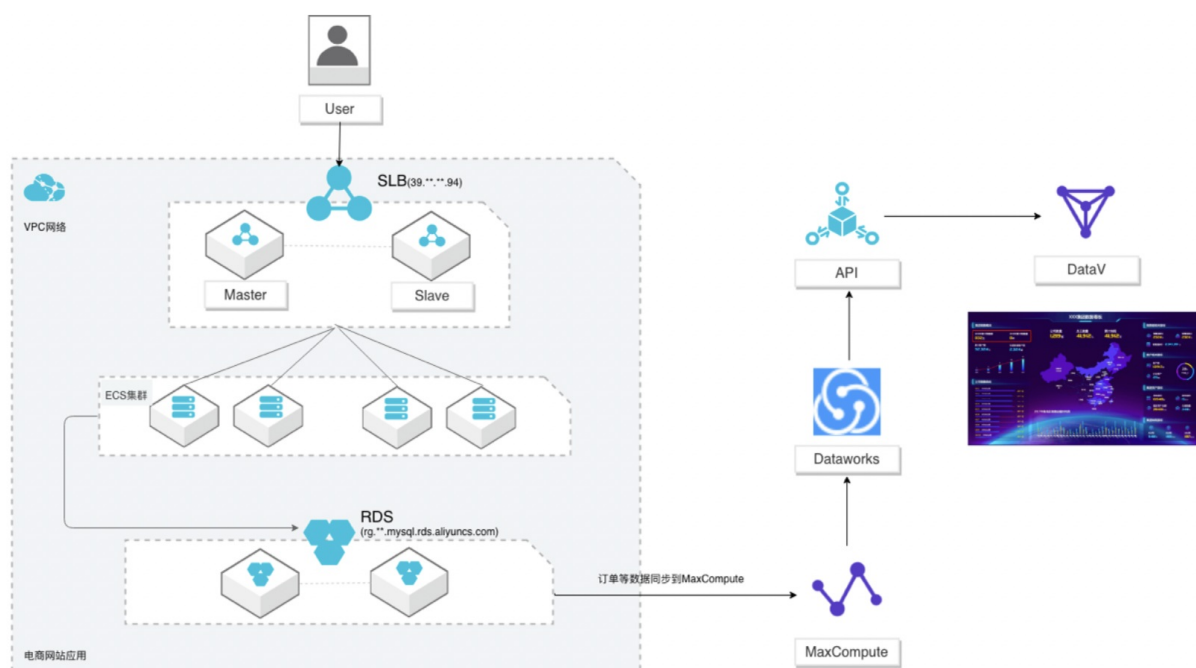
场景描述

电商网站销售数据通过大数据分析后将业务指标数据在大屏幕上展示，如销售指标、客户指标、销售排名、订单地区分布等。大屏上销售数据可视化动态展示，效果震撼；触控大屏支持用户自助查询数据，极大地增强数据的可读性。

解决问题

- 电商网站数据看板
- 全国/全球业务态势分析
- 互联网金融风险数据监控

部署架构图



24.基于MaxCompute的大数据BI分析

在互联网、电商及游戏等行业通常需要对海量数据做快速实时分析和决策。本文演示如何将业务数据和日志数据通过MaxCompute处理后汇总到ADB，并通过QuickBI等工具进行可视化分析的方案。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践](#)

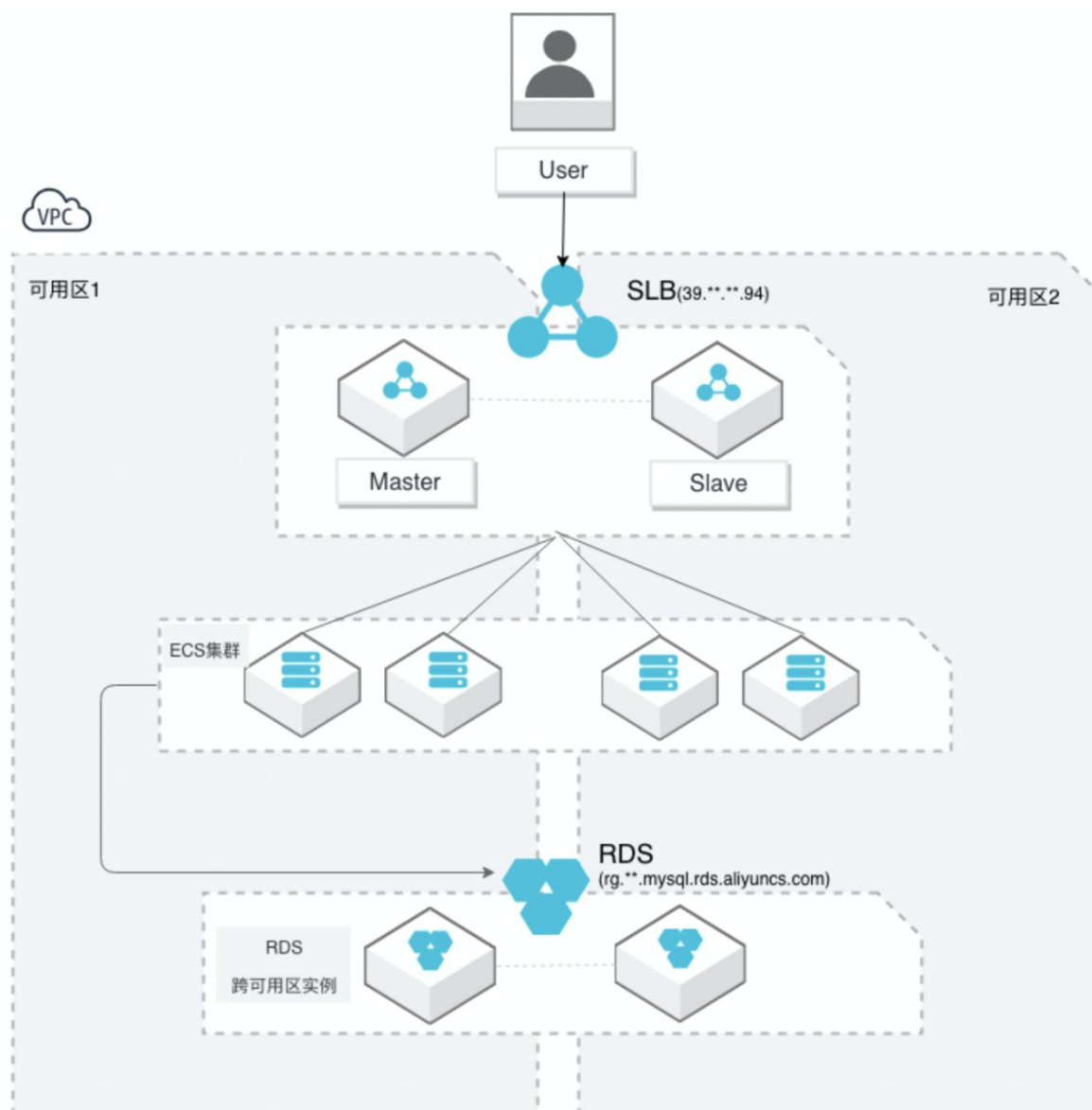
场景描述

本文以电商行业为例，将业务数据和日志数据使用MaxCompute做ETL之后，同步到分析型数据库进行实时分析，之后通过Quick BI进行快速可视化展示。

解决问题

- 互联网行业、电商、游戏行业等网站、App、小程序应用内BI分析场景
- 可扩展到各类网站BI分析场景使用

部署架构图



25.EMR本地盘实例大规模数据集测试

使用阿里云EMR和D1本地盘实例实现大数据分析集群架构以及大规模数据集测试。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践](#)

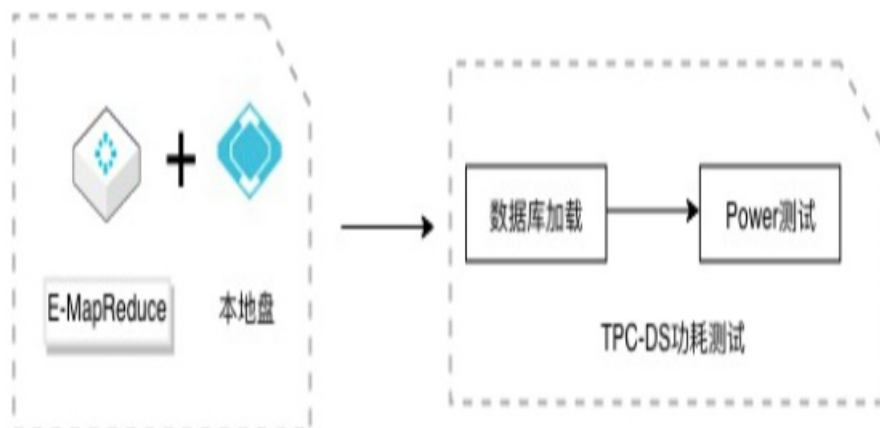
场景描述

阿里云为了满足大数据场景下的存储需求，在云上推出了本地盘D1机型，这个系列提供了本地盘而非云盘作为存储，提高了磁盘的吞吐能力，发挥Hadoop的就近计算优势。阿里云EMR产品针对本地盘机型，推出了一整套的自动化运维方案，帮助用户方便可靠地使用本地盘机型，不需要关注整个运维过程，同时保证数据的高可靠和服务的高可用。

解决问题

- 云盘多份冗余数据导致成本高
- 磁盘吞吐量不高
- 节点的高可靠分布问题
- 本地盘与节点的故障监控问题
- 数据迁移时自动决策问题
- 自动故障节点迁移与数据平衡问题

部署架构图



26.EMR弹性低成本离线大数据分析

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多最佳实践](#)

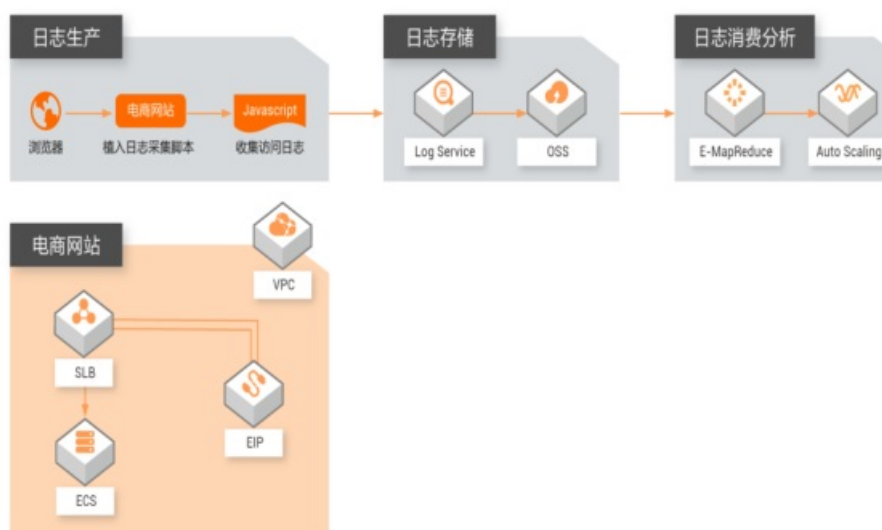
场景描述

基于阿里云的E-MapReduce(EMR)、对象存储OSS、日志服务SLS、抢占式ECS实例构建弹性、低成本的计算与存储分离架构的海量离线大数据分析日志分析系统。

解决问题

- 大数据平台运维管理成本高
- 计算资源弹性能力不足
- 海量日志数据下的存储容量不足
- 离线分析场景下的计算资源成本控制
-

方案架构



27.互联网、电商及游戏行业实时BI分析

在互联网、电商及游戏等行业通常需要对海量数据做快速实时分析和决策，本文演示如何将业务数据和日志数据汇总到ADB，并通过QuickBI等工具进行可视化分析的方案。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践](#)

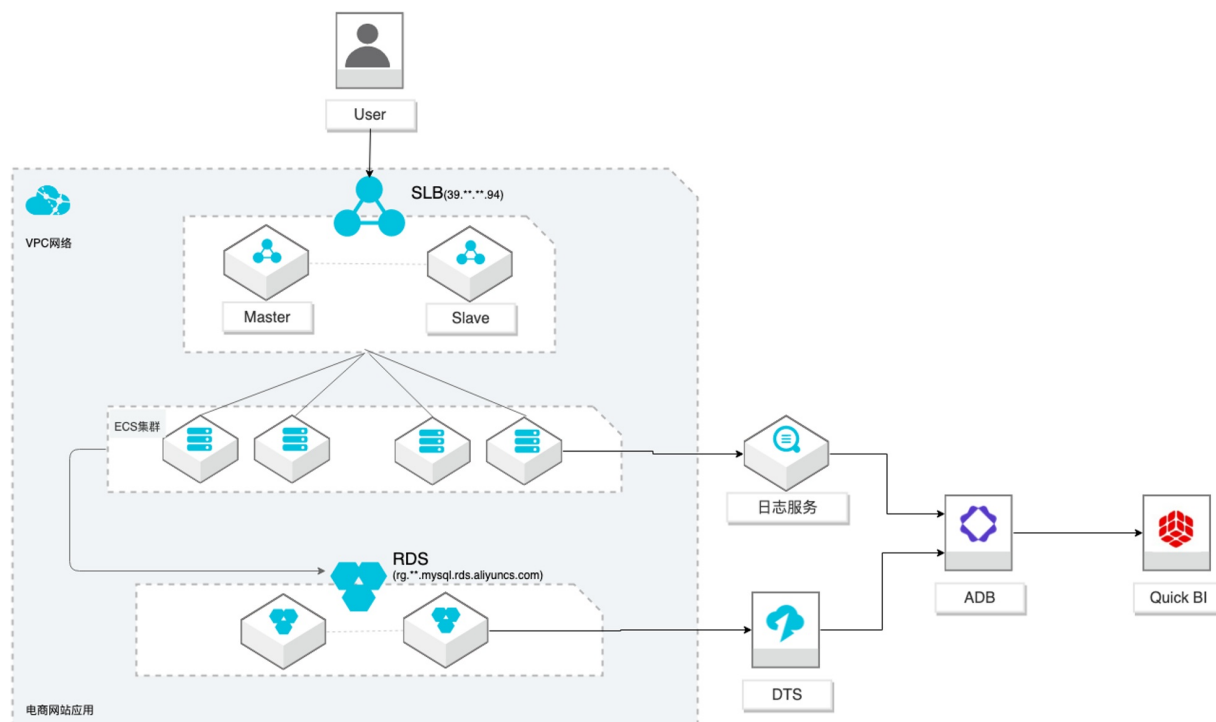
场景描述

本文以电商行业为例，将业务数据和日志数据同步到分析型数据库，之后通过Quick BI做实时可视化分析。相对于传统的关系型数据库，阿里云分析型数据库MySQL版只需要几毫秒的时间，即可查询PB级数据并从中找到匹配信息。

解决问题

- 互联网行业、电商、游戏行业等网站、App、小程序应用内BI分析场景
- 在线运营、运营指标实时化分析等场景
- 可扩展到各类网站BI分析场景使用

部署架构图



28.低成本基于抢占式ECS实例进行离线大数据分析

使用阿里云抢占式ECS实例以及OSS等云上产品构建低成本弹性大数据分析系统。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践](#)

场景描述

基于阿里云的抢占式ECS实例以及弹性伸缩能力，同时结合阿里云对象存储OSS服务搭建离线大数据分析集群，实现计算与存储分离架构，同时降低计算资源成本和提升计算资源利用率，且拥有海量存储能力。

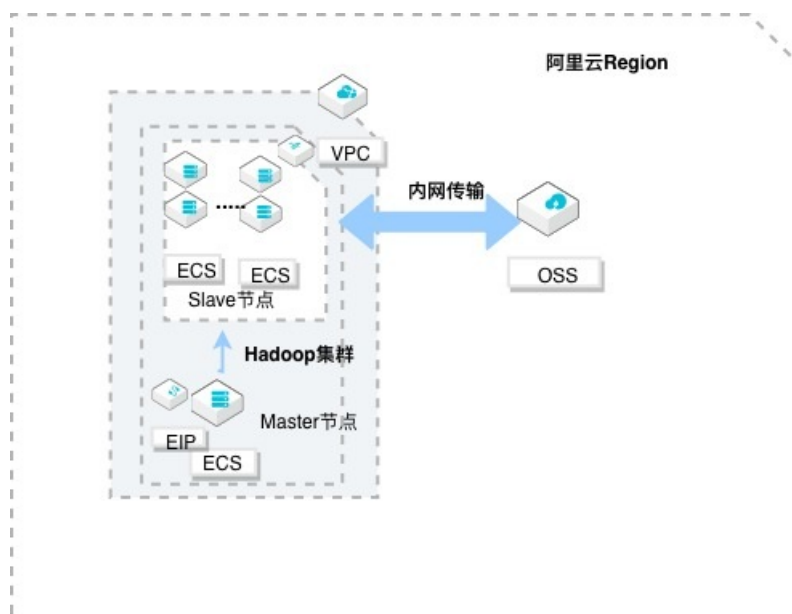
方案优势

- 计算与存储资源成本优化
- 大数据分析集群海量存储能力
- 计算资源弹性伸缩
- 运维简便
- 计算与存储分离架构，集群高性能

解决问题

- 离线大数据分析集群计算和存储资源成本管控困难
- 计算资源弹性能力不足
- 大数据量大导致存储容量不足
- 大数据分析集群云资源性能优化

部署架构图



29.DTS数据同步集成MaxCompute数仓

介绍通过DTS实现从RDS到MaxCompute的数据同步集成，并介绍如何使用DTS和MaxCompute联合实现数据。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践。](#)

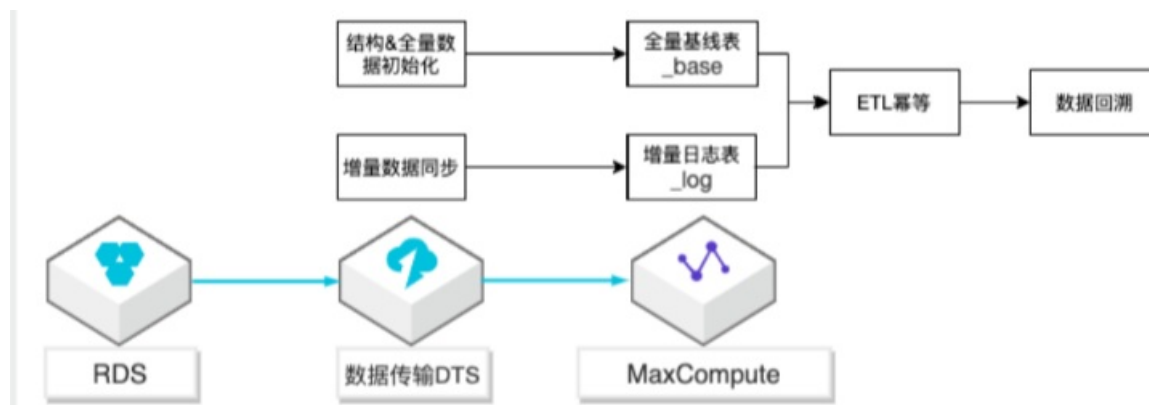
场景描述

本文Step by Step介绍了通过数据传输服务DTS，实现从云数据库RDS到MaxCompute的数据同步集成，并介绍如何使用DTS和MaxCompute数仓联合实现数据ETL幂等和数据生命周期快速回溯。

解决问题

- 实现大数据实时同步集成
- 实现数据ETL幂等
- 实现数据生命周期快速回溯

部署架构图



30.Spark on ECI大数据分析

Spark和云原生结合提供计算与存储分离的高性价比大数据分析。

直达最佳实践

[点击查看最佳实践详情](#)

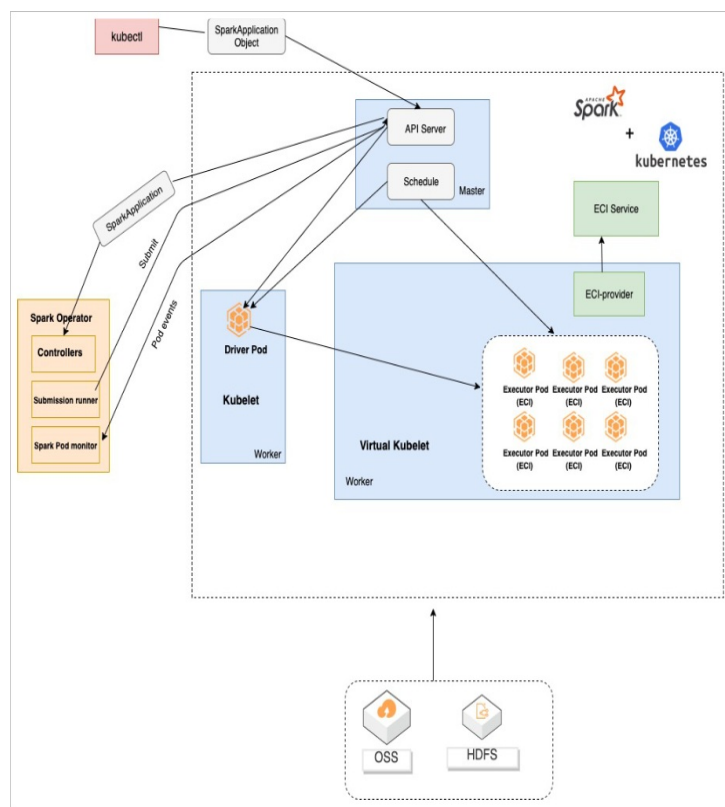
更多最佳实践

[点击查看更多阿里云最佳实践](#)

场景描述

Spark作为快速、通用的大规模数据处理平台，更多关注Spark Application的管理，底层实际资源调度和管理更多的是依靠外部平台的支持例如Mesos、YARN、Kubernetes等。借助阿里云的容器服务Kubernetes版（ACK）、弹性容器组实例（ECI）、文件存储HDFS或者对象存储OSS提供灵活弹性计算资源弹性可扩展、计算与存储分离架构、成本可控的Spark on ECI解决方案实践。

解决架构



方案优势

- 计算引擎弹性扩缩容，兼顾资源弹性与计算资源成本优化。
- 计算与存储分离架构，结合阿里云原生云存储产品，海量数据湖优势。
- Kubernetes原生的调度性能优势，提升在大规模分析作业时的分析性能优势分。
- 集群资源隔离和按需分配。

解决问题

- 计算资源弹性能力不足，计算资源成本管控能力欠缺。
- 集群资源调度能力和隔离能力不足。
- 计算与存储无法分离，大数据量分析时出现数据存储资源瓶颈。
- Spark submit方式提交分析作业参数支持有限等缺点。

31.数据湖：在线学习场景数据分析

以在线教育中答题闯关类应用为例，使用Web Server模拟这类日志数据的分析处理。

直达最佳实践

[点击查看最佳实践详情](#)

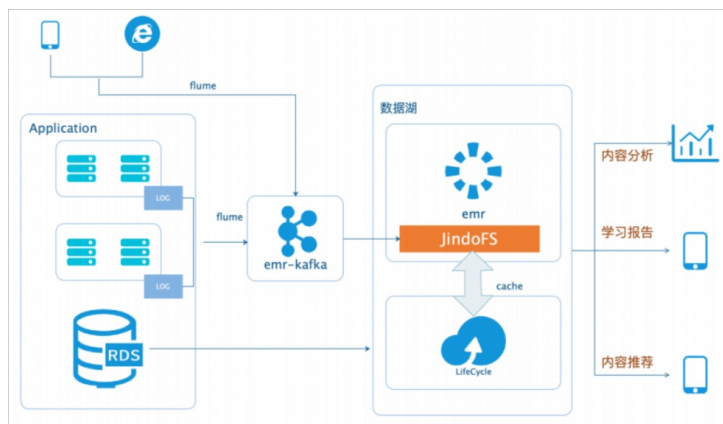
更多最佳实践

[点击查看更多阿里云最佳实践](#)

场景描述

本场景以在线教育中一个答题闯关类的应用为例，使用Web Server来模拟演示这类日志数据的分析处理。通过Nginx和Python flask搭建Web Server，模拟应用中的关键页面，比如登录、课程内容等，之后构造若干用户使用的模拟日志数据，投递到数据湖进行分析后获取应用 PV、UV、课程内容访问排行、平均得分等等。

解决架构



解决问题

- 基于数据湖（EMR+OSS）搭建大数据平台
- EMR和OSS使用和配置
- 数据统一存储到OSS

32.互联网、游戏等行业实时数仓构建

基于Flink构建实时数仓：采集日志、拆分日志、处理写入ADB存储和HDFS。

直达最佳实践

[点击查看最佳实践详情](#)

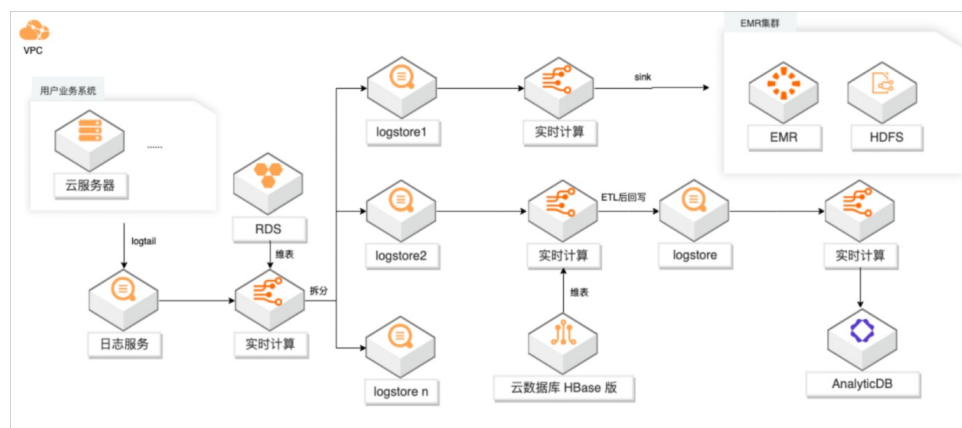
更多最佳实践

[点击查看更多阿里云最佳实践](#)

场景描述

互联网、游戏等行业客户基于Flink构建实时数仓，通过Logtail采集日志，使用Flink实现对日志的拆分、处理，根据业务需求写入不同的数据存储，实时分析使用ADB存储、离线分析使用HDFS存储，构建高效、稳健的实时数据服务。

解决架构



解决问题

- 日志采集、处理及分析
- 日志使用Flink拆分到不同logstore
- 日志使用Flink实时写入HDFS
- 日志数据实时ETL
- 日志实时写入ADB

33.基于Flink的咨询场景实时数仓

针对资讯聚合类业务场景，Step by Step介绍如何搭建实时数仓。

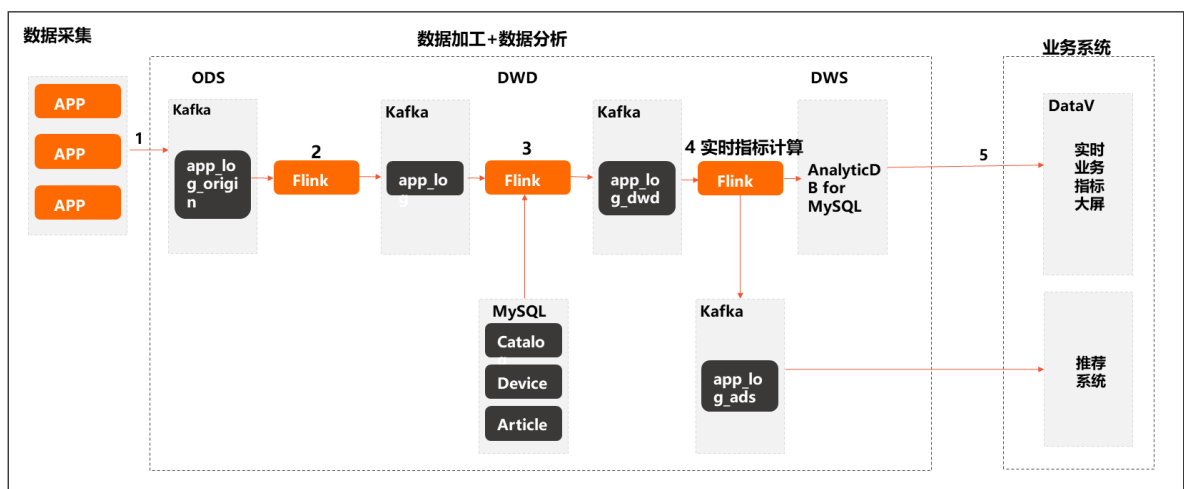
直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践](#)

解决架构



解决问题

- 如何搭建实时数仓。
- 通过实时计算Flink实现实时ETL和数据流。
- 通过实时计算Flink实现实时数据分析。
- 通过实时计算Flink实现事件触发。

34.基于Dataworks的大数据一站式开发及数据治理

基于Dataworks做大数据一站式开发，包含数据实时采集到kafka通过实时计算对数据进行ETL写入HDFS，使用Hive进行分析。通过Dataworks进行数据治理，数据地图查看数据信息和血缘关系，数据质量监控异常和报警。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践](#)

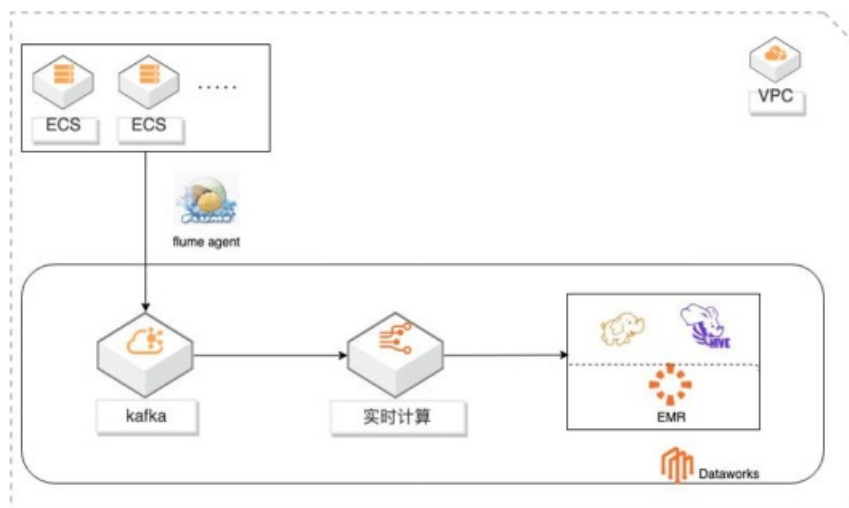
解决问题

- 日志采集、处理及分析
- 日志使用Flink实时写入HDFS
- 日志数据实时ETL
- 日志HIVE分析
- 基于dataworks一站式开发
- 数据治理

方案优势

- 大数据一站式开发，完善的数据治理能力。
- 性能优越：高吞吐，高扩展性。
- 安全稳定：Exactly-Once，故障自动恢复，资源隔离。
- 简单易用：SQL语言，在线开发，全面支持UDX。
- 功能强大：支持SQL进行实时及离线数据清洗、数据分析、数据同步、异构数据源计算等DataLake相关功能，以及各种流式及静态数据源关联查询。
- 安全：原生的多租户系统，以项目进行隔离，所有计算任务在安全沙箱中运行。

部署架构图



相关产品

- 文件存储HDFS

阿里云文件存储HDFS（Apsara File Storage for HDFS）提供标准的HDFS访问协议，用户无需对现有大数据分析应用做任何修改，即可使用具备无限容量及性能扩展、单一命名空间、高可靠和高可用等特性的分布式文件系统。

更多关于文件存储HDFS的介绍，参见[文件存储HDFS产品详情页](#)。

- 实时计算

实时计算（AlibabaCloudRealtimeCompute，PoweredbyVerverica）是阿里云提供的基于ApacheFlink构建的企业级大数据计算平台。在PB级别的数据集上可以支持亚秒级别的处理延时，赋能用户标准实时数据处理流程和行业解决方案；支持DatastreamAPI作业开发，提供了批流统一的FlinkSQL，简化BI场景下的开发；可与用户已使用的大数据组件无缝对接，更多增值特性助力企业实时化转型。

更多关于实时计算的介绍，参见[实时计算产品详情页](#)。

- E-MapReduce

阿里云E-MapReduce (EMR) 是构建在阿里云云服务器ECS上的开源Hadoop、Spark、HBase、Hive、Flink生态大数据PaaS产品。提供用户在云上使用开源技术建设数据仓库、离线批处理、在线流式处理、即时查询、机器学习等场景下的大数据解决方案。

更多关于E-MapReduce的介绍，参见[E-MapReduce产品详情页](#)。

- DataWorks

DataWorks基于MaxCompute/EMR/MC-Hologres等大数据计算引擎，为客户提供专业高效、安全可靠的一站式大数据开发与治理平台，自带阿里巴巴数据中台与数据治理最佳实践，赋能各行业数字化转型。每天阿里巴巴集团内部有数万名数据/算法工程师正在使用DataWorks，承担集团99%数据业务构建。

更多关于DataWorks的介绍，参见[DataWorks产品详情页](#)。

35.利用交互分析（Hologres）进行数据查询解决方案

数据经由实时大数据与历史数据聚合分析后通过BI工具展示。该方案打通实时和离线数据的壁垒，并将数据实时展示在BI工具中；针对IoT、金融、广告等领域需要实时数据和离线数据结合使用的领域有所帮助。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践](#)

场景描述

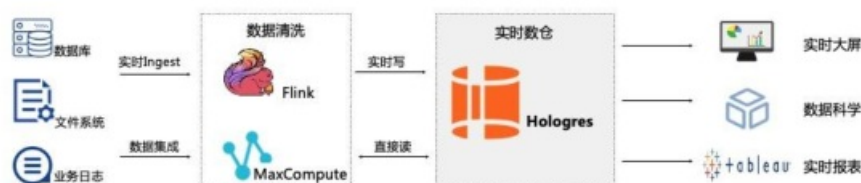
随着收集数据的方式不断丰富，企业信息化程度越来越高，企业掌握的数据量呈TB、PB或EB级别增长。同时，数据中台的快速推进，使数据应用主要为数据支撑、用户画像、实时圈人及广告精准投放等核心业务服务。高可靠和低延时地数据服务成为企业数字化转型的关键。

Hologres致力于低成本和高性能地大规模计算型存储和强大的查询能力，为您提供海量数据的实时数据仓库解决方案和实时交互式查询服务。

解决问题

- 加速查询MaxCompute数据
- 快速搭建实时数据仓库
- 无缝对接主流BI工具。

部署架构图



相关产品

• MaxCompute 交互式分析（Hologres）

MaxCompute 交互式分析（Hologres）是为大数据设计的实时交互式分析产品，它与MaxCompute无缝打通，支持数据实时写入，支持PB级数据进行高并发、低延时的分析处理，兼容PostgreSQL协议，可以使用您最熟悉的BI工具对海量数据进行自助的多维分析透视和业务探索，同时也支持超高QPS点查能力，满足数仓分析、服务一体化需求。

更多关于MaxCompute 交互式分析（Hologres）的介绍，参见[MaxCompute 交互式分析（Hologres）产品详情页](#)。

- 大数据计算服务MaxCompute

大数据计算服务（MaxCompute，原名ODPS）是一种快速、完全托管的EB级数据仓库解决方案。随着数据收集手段不断丰富，行业数据大量积累，数据规模已增长到了传统软件行业无法承载的海量数据（百TB、PB、EB）级别。MaxCompute致力于批量结构化数据的存储和计算，提供海量数据仓库的解决方案及分析建模服务。由于单台服务器的处理能力有限，海量数据的分析需要分布式的计算模型。分布式的计算模型对数据分析人员要求较高且不易维护。数据分析人员不仅需要了解业务需求，同时还需要熟悉底层分布式计算模型。MaxCompute为您提供完善的数据导入方案以及多种经典的分布式计算模型，您可以不必关心分布式计算和维护细节，便可轻松完成大数据分析。

更多关于大数据计算服务MaxCompute的介绍，参见[大数据计算服务MaxCompute产品文档](#)。

- 实时计算Flink

阿里云实时计算Flink版（Alibaba Cloud Realtime Compute for Apache Flink，Powered by Ververica）是阿里云基于Apache Flink构建的企业级、高性能实时大数据处理系统，由Apache Flink创始团队官方出品，拥有全球统一商业化品牌，完全兼容开源Flink API，提供丰富的企业级增值功能。

更多关于实时计算Flink的介绍，参见[实时计算Flink产品文档](#)。

详细信息

[点击查看最佳实践详情](#)。

更多最佳实践

[点击查看更多阿里云最佳实践](#)。

36.MaxCompute湖仓一体方案

自建数据湖与云数仓的融合解决方案，将MaxCompute与自建的Hive集群做数据打通，通过存储共享，元数据镜像等方式，解决传统模式下的存储冗余，计算资源弹性能力弱的痛点。可大幅度增强系统的资源弹性，解决业务高峰期计算资源不足的问题。

直达最佳实践

[点击查看最佳实践详情](#)

更多最佳实践

[点击查看更多阿里云最佳实践](#)

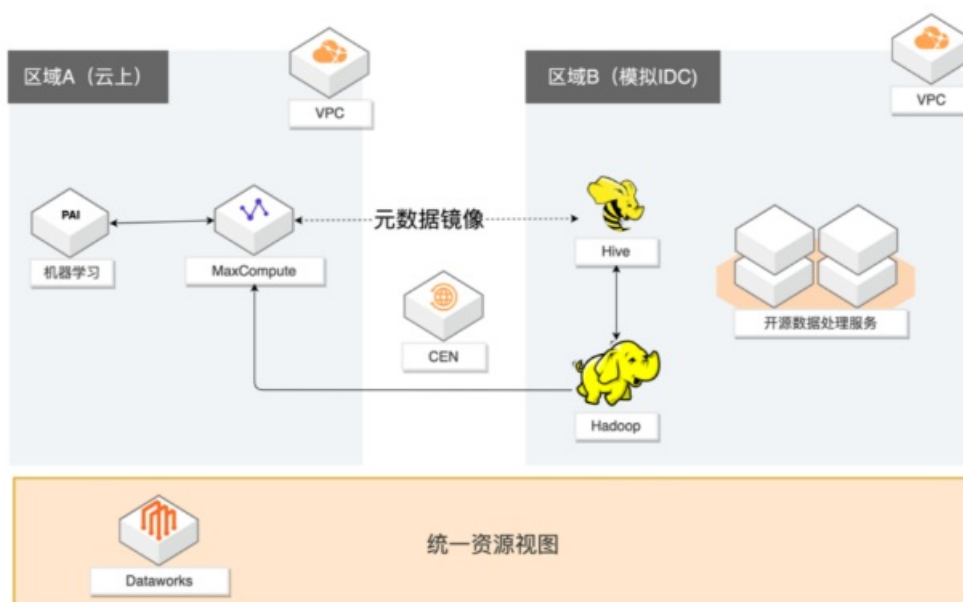
方案概述

客户在IDC或者公有云环境基于Hadoop自建数据湖方案，数据集中保存在HDFS文件系统用于数据分析任务。当线下资源无法满足海量的计算需求后，通过打通MaxCompute与线下数据湖之间的网络与元数据，MaxCompute可直接使用存储在自建数据湖中的数据进行计算，大幅减少集群的计算峰值压力。本实践将介绍全链路打通MaxCompute与自建数据湖的方案。

适用场景

- 自建数据湖有较大弹性计算需求
- 自建数据湖的数据治理能力优化
- SQL类计算效率提升和成本控制
- 构建混合云数据处理方案

技术架构



方案优势

- 性能优越：高吞吐，高扩展性。

- 无侵入性：现有业务无需改造。
- 性能优化：MaxCompute在SQL上做了大量优化与能力沉淀，可提高SQL运行性能，降低计算成本。
- 灵活管理：元数据实时同步。
- 资源弹性：利用MaxCompute计算池弹性进行海量数据计算。