DTCC 2010-2022

第十三届中国数据库技术大会

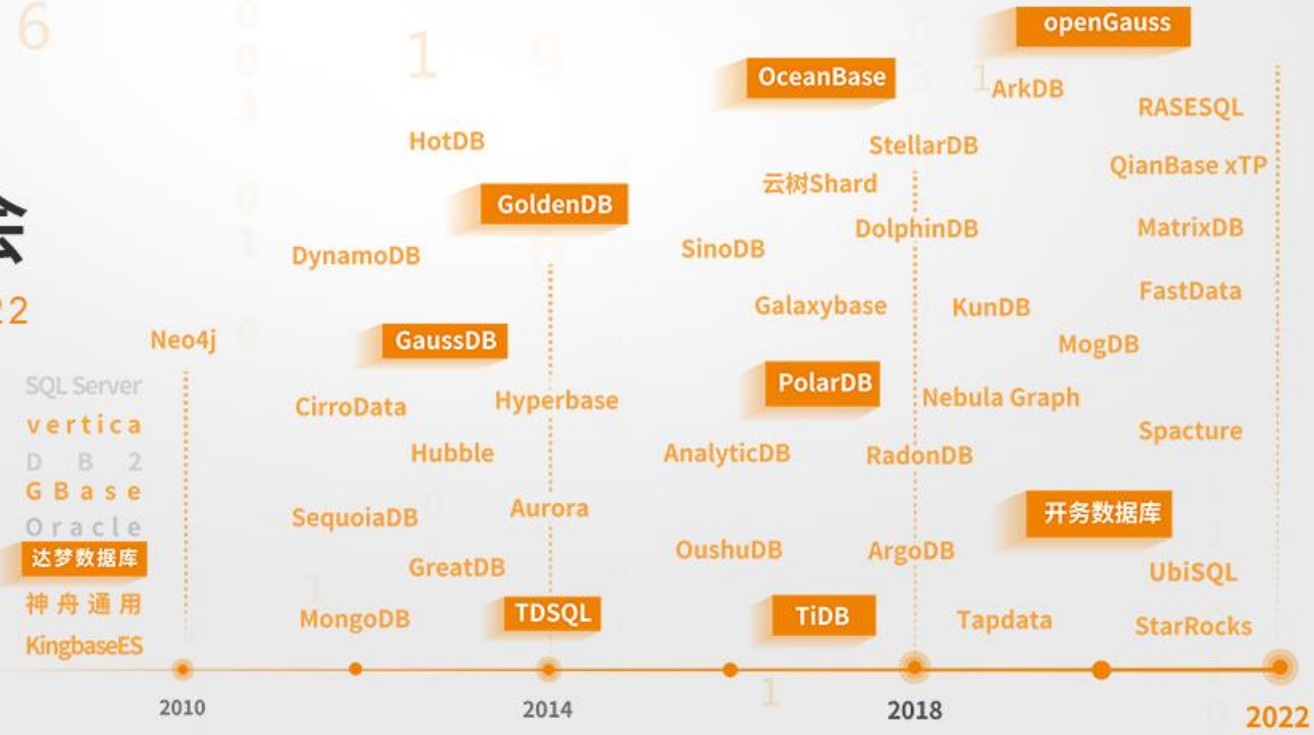
DATABASE TECHNOLOGY CONFERENCE CHINA 2022

数据智能 价值创新

IT168.com ChinaUnix.net ITPUB

线上直播 | 2022/12/14-16

数据来源：数据库产品上市商用时间



云原生无服务器数仓最佳实践与实时数仓架构

潘超
亚马逊云科技
数据分析专家

议题

1. 亚马逊云科技云原生数仓Redshift 10年架构演进
2. Redshift Serverless架构设计与应用场景
3. 基于Redshift的云原生实时数仓架构与最佳实践

议题

1. 亚马逊云科技云原生数仓Redshift 10年架构演进
2. Redshift Serverless架构设计与应用场景
3. 基于Redshift的云原生实时数仓架构与最佳实践

全球数万客户每天使用Redshift来进行数据分析

Amazon Redshift | 最受欢迎的云原生数仓

游戏

金融

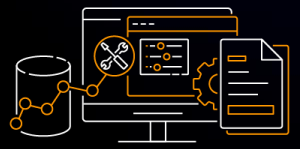
医疗健康

消费娱乐

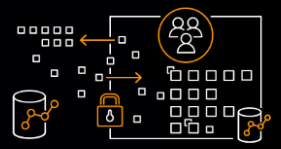
互联网

亚马逊云科技

典型数仓场景



业务运营与商业智能

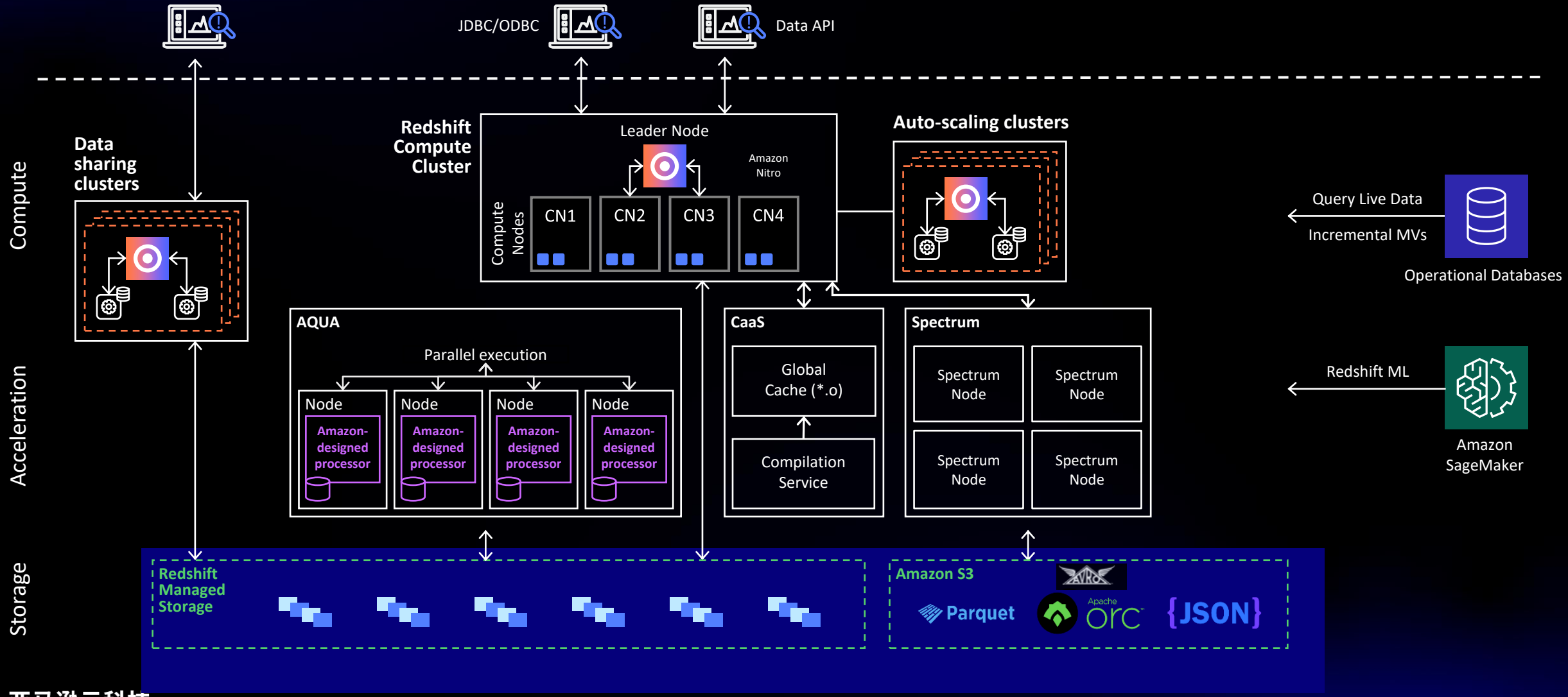


实时数仓分析

查询、报表与数据分析

机器学习与分析预测

Redshift的架构演进



Amazon Redshift的创新更新

简单易用的分析

秒级从数据获得洞察，
无需关心基础架构

新功能!

无服务器

新功能!

查询编辑器v2

更新!

自动化数据仓库管理

新功能!

自动物化视图

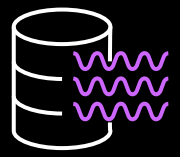

数据API


Amazon Redshift顾问


CloudFormation模板

新功能!

Grafana插件



分析所有数据

在运营数据库、数据湖和数据仓库中进行复杂、可扩展的实时和预测分析

更新!

数据共享

新功能!

数据交换集成

更新!

Redshift ML

更新!

联邦查询

更新!

地理空间增强


SUPER数据类型

卓越性价比

与其他云数据仓库相比，提供超过3X的性价比，并可动态扩展以提高复杂和关键工作负载的查询速度


RA3节点和托管存储


AQUA

新功能!

写入的并发缩放

更新!

SQL增强和迁移支持


安全、治理&合规


工作负载管理增强功能

亚马逊云科技

议题

1. 亚马逊云科技云原生数仓Redshift 10年架构演进
2. **Redshift Serverless架构设计与应用场景**
3. 基于Redshift的云原生实时数仓架构与最佳实践

Amazon Redshift
无服务器架构

自动、智能、按需扩展计算与存储资源

数据查询与分析

高性能与高并发

按需计费，优化成本

客户

只需关心数据价值探索

自动扩展

计算资源自动分配

自动升级

自动冗余

实时监控

自动备份

无宕机运维

安全与加密

亚马逊云科技

其它所有基础架构管理

灵活的云原生数仓部署模式

Amazon Redshift
集群模式

细粒度控制与定制化集群

Amazon Redshift
无服务器

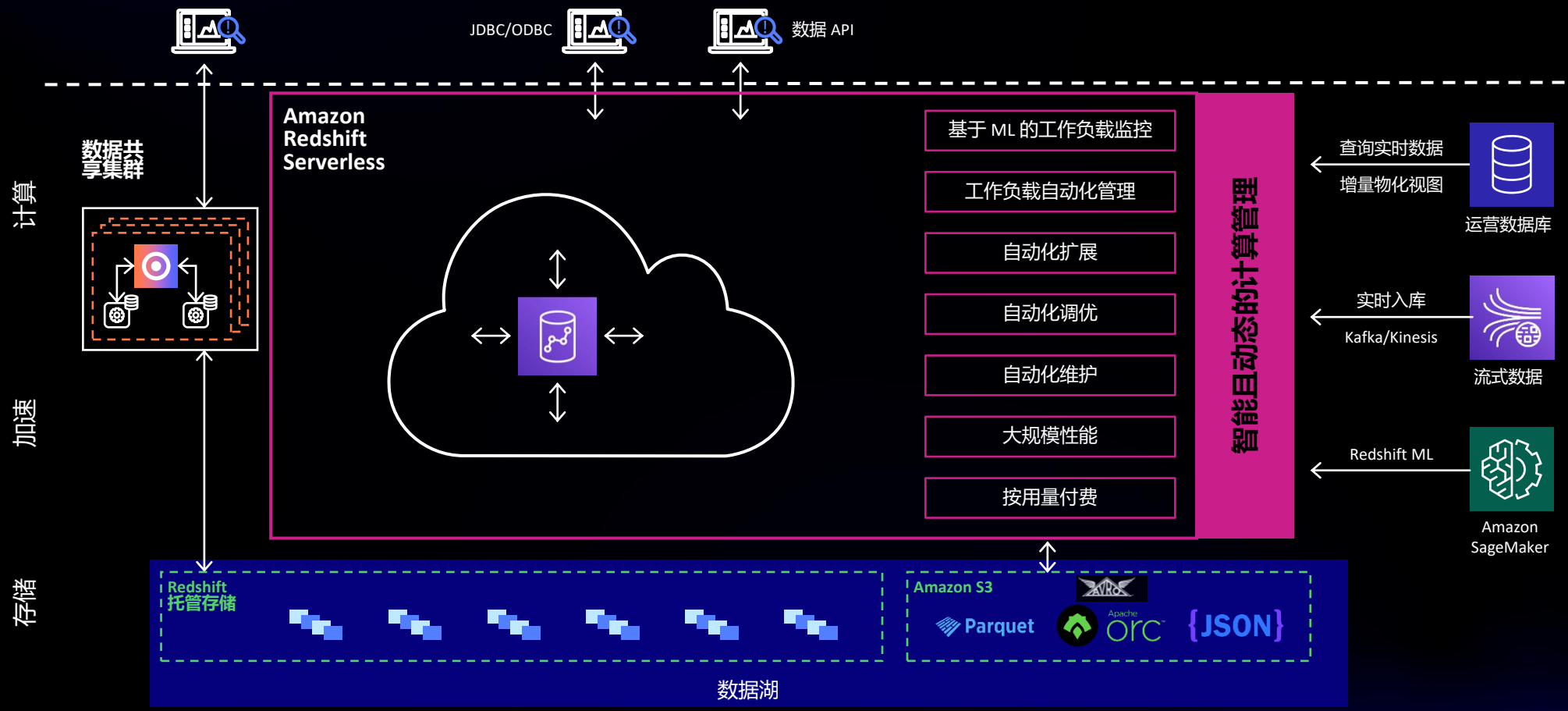
无需管理底层基础架构

亚马逊云科技 *来帮助管理云基础设施*

亚马逊云科技

Redshift Serverless架构

2022



亚马逊云科技

Redshift Serverless特点

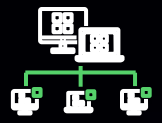
简化用户体验

自动扩展资源，而无需管理数据仓库集群

智能动态计算

自动调配和扩展数据仓库容量，提供一致快速的用户体验

支持所有Amazon Redshift的功能和性能特性



利用Amazon Redshift丰富的SQL功能、无缝数据湖集成和业界领先的性价比

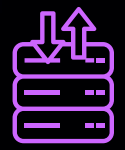
按需付费

按秒使用的计算能力和持续时间来计费，空闲时不收费

统一计费

计算

- 新的标准化计算单元 – Redshift Processing Unit (RPU)
- 按照 RPU 使用小时数计费，计量精确至秒
- 基础的数据仓库、扩展容量和数据湖查询也包含在相同的 RPU 小时数之中
- 不含并发扩展和 Spectrum 费用

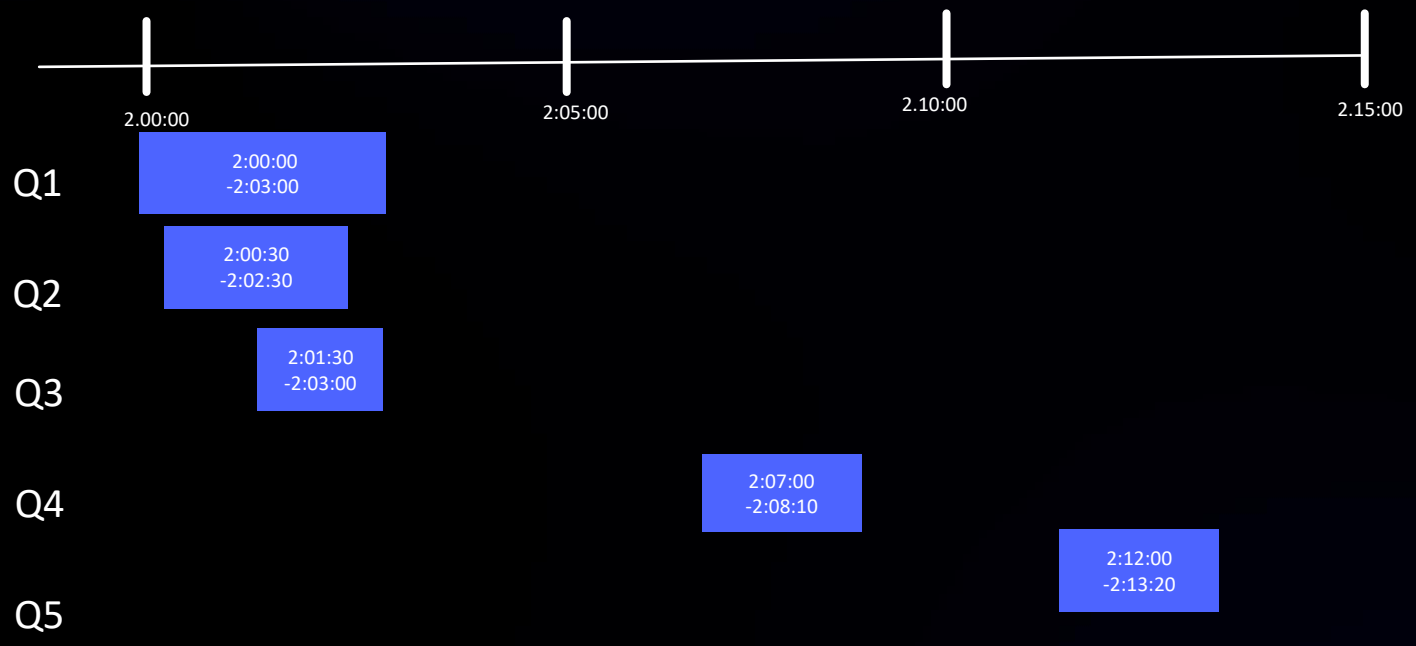


存储

- Redshift 托管存储和用户快照按照固定的每 GB 月费率收取
- 可免费将数据仓库还原为过去 24 小时内以 30 分钟为间隔的时点

按用量付费

只为工作负载运行期间使用的计算容量付费（按秒计费）



计费时段	查询执行时间
@2:03	3 分钟（Q1、Q2、Q3）
@2:09	1 分 10 秒（Q4）
@2:14	1 分 20 秒（Q5）
总时长	5 分 30 秒

闲置时段不收费

亚马逊云科技

Redshift Serverless适用多种应用场景

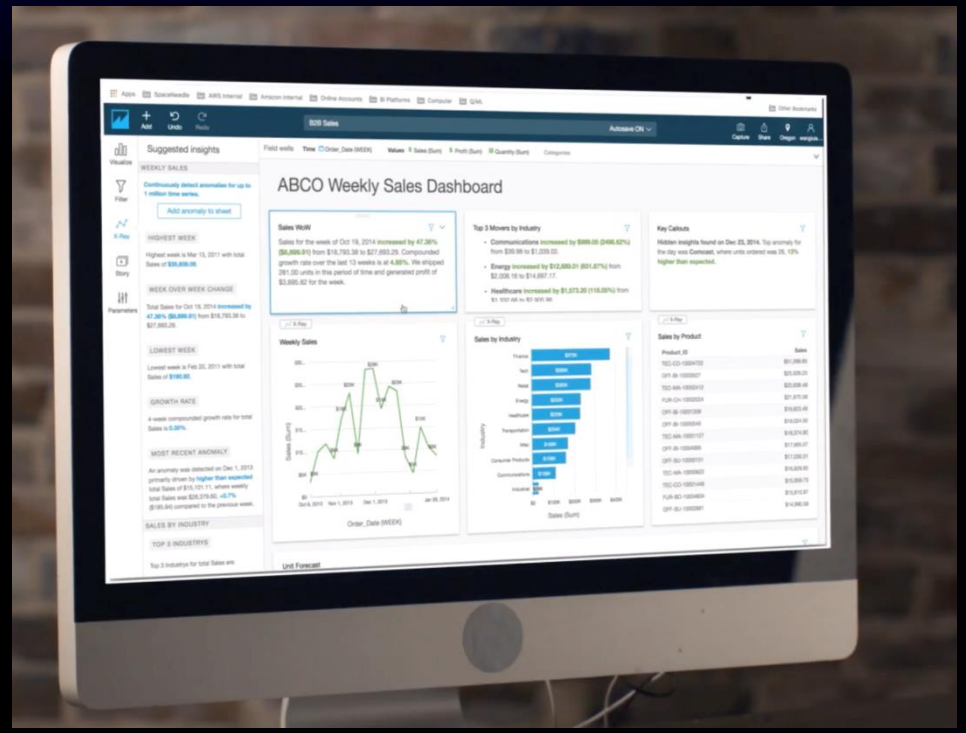
自助分析服务

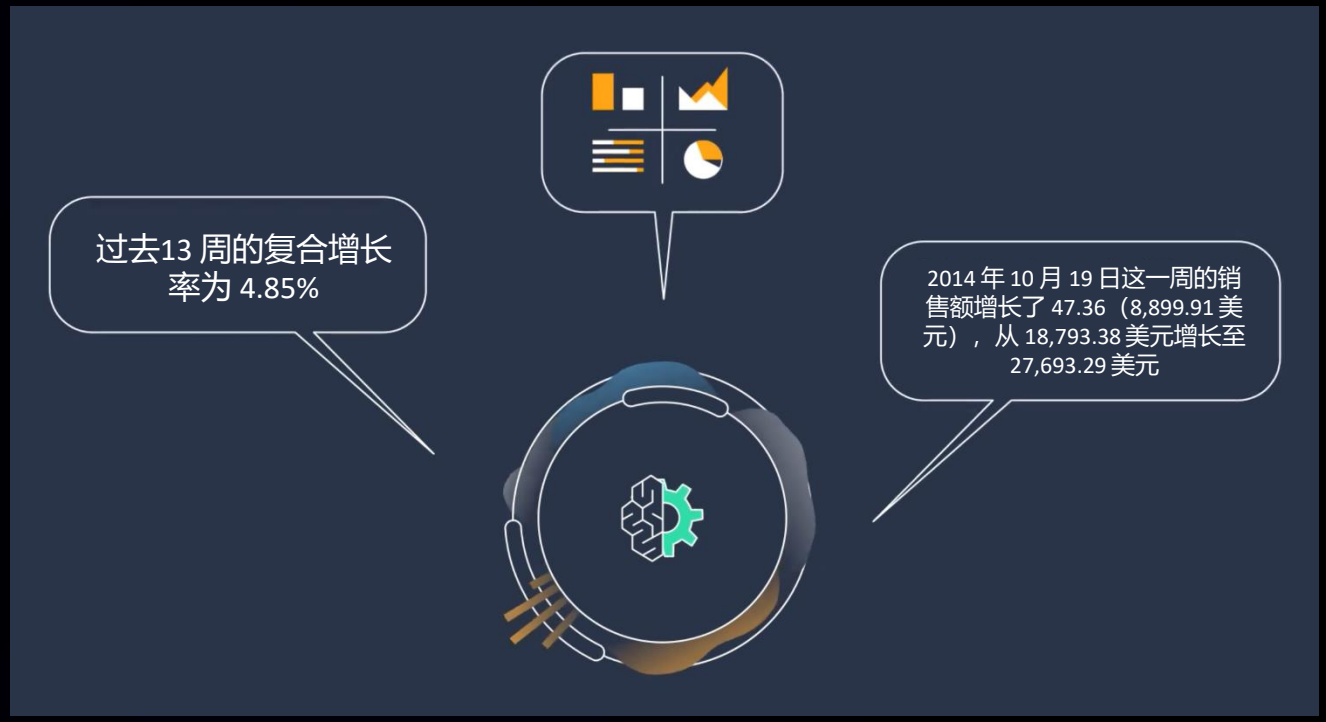
- 机器学习
- 报表与仪表盘
- 实时分析
- 无数据移动的数据共享

自动扩展

- 应对从小到大的ad-hoc资源需求
- 应对无法预测或瞬态高峰的资源负载需求
- 有规律的高、低负载窗口
- 端到端无服务器架构 – 数据库、分析、机器学习

Serverless应用场景一：简单易用、轻松分析





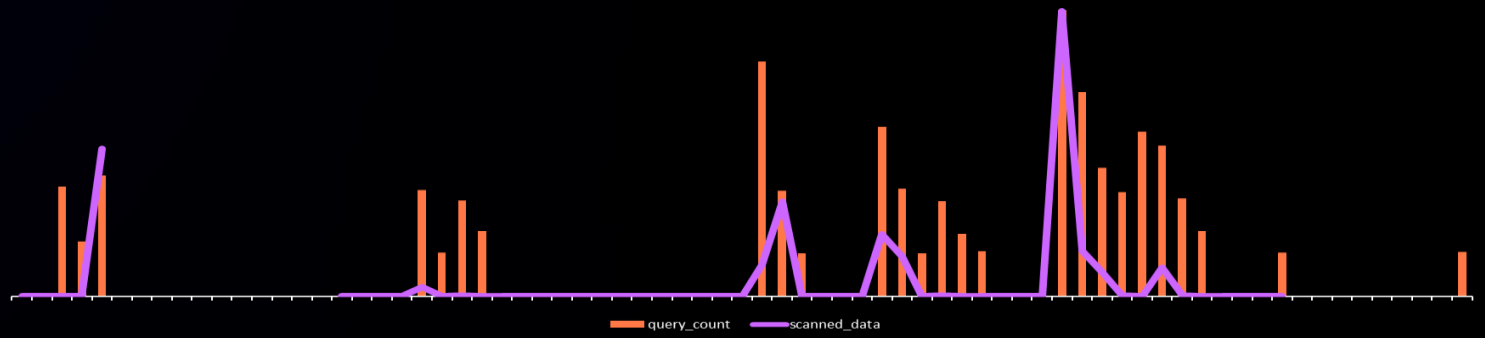
原因

- 无需考虑基础设施即可轻松上手
- 无需选择节点类型和数量并创建和管理集群
- 用例范例：开发/测试环境以及即席业务分析
- 按用量付费

亚马逊云科技

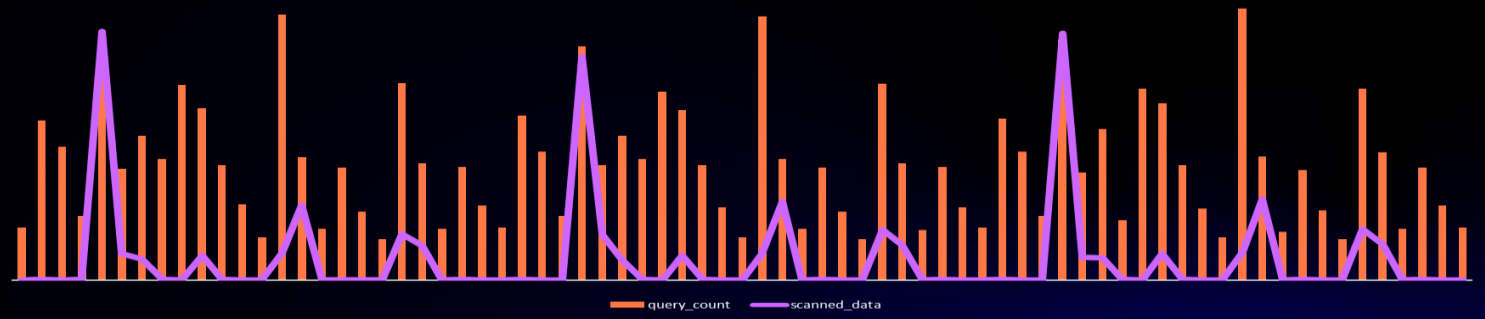
Serverless应用场景二：支持多种工作负载

多样化工作负载



周期性工作负载

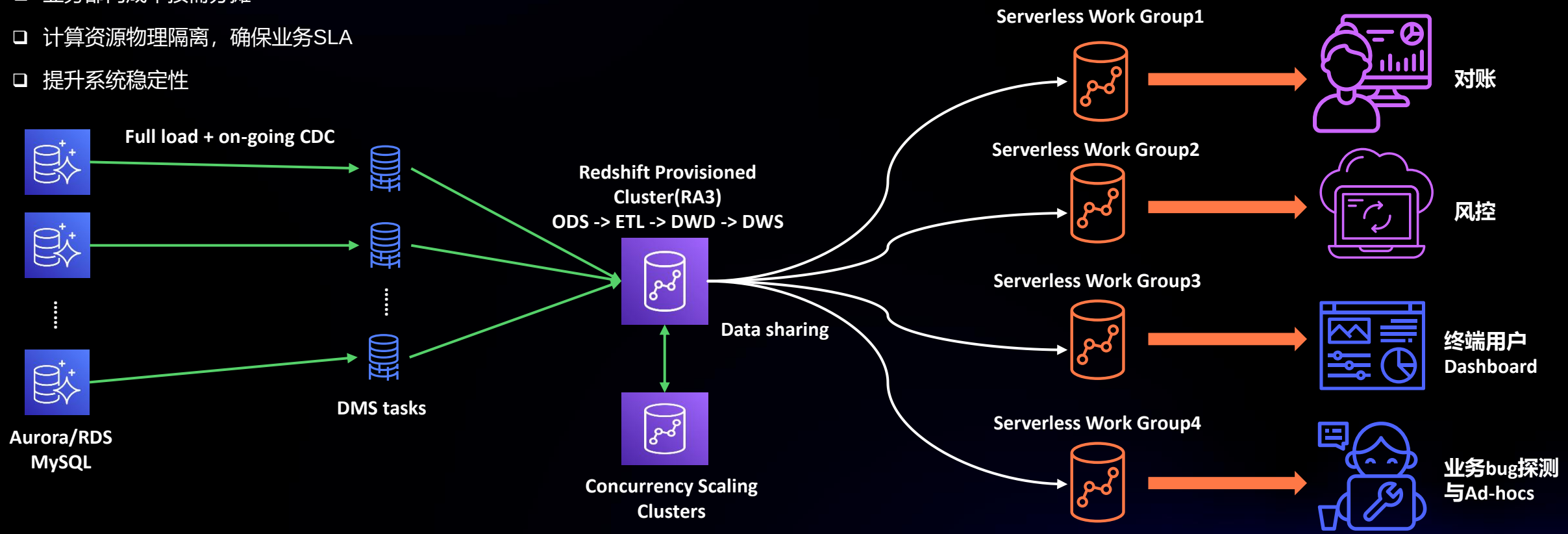
稳态中存在峰值



亚马逊云科技

Serverless应用场景三：Provisioned与Serverless共存

- 通过DMS将数据实时CDC到Redshift Provisioned Cluster
- 在Provisioned Cluster里进行数据加工、转换，构建DWD/DWS层
- 利用Data Sharing功能将数据共享给不同的Serverless WorkGroup
  - 业务部门成本按需分摊
  - 计算资源物理隔离，确保业务SLA
  - 提升系统稳定性



亚马逊云科技

议题

1. 亚马逊云科技云原生数仓Redshift 10年架构演进
2. Redshift Serverless架构设计与应用场景
3. **基于Redshift的云原生实时数仓架构与最佳实践**

客户对实时数仓的需求

数据摄入高吞吐、低延迟

处理任意数量的流式数据并处理来自数十万个数据源的数据，同时提供非常低的延迟和高带宽，可以实现可以在几秒钟而不是几分钟内获得数据分析见解。

简单配置与使用

在几秒钟内实现实时分析，无需管理复杂的管道，完全托管，支持流处理应用，无需基础设施管理。

提高生产力

能使用SQL对流数据进行丰富的分析，无需依赖其它语言。

实时数仓应用场景

改善游戏体验

通过分析玩家的实时数据，专注于游戏转化率、玩家留存和优化游戏体验。

在线广告用户点击流数据分析

客户通常在一次会话中访问数十个网站，但营销人员一般只分析自己的网站。通过将数据实时摄入到仓库中，可以实时评估您的客户足迹和行为。

零售行业实时销售分析

近实时访问和可视化所有POS零售销售交易数据，以实现实时分析、报告和可视化。

实时应用洞察

通过访问和分析应用程序日志文件和网络日志中的流数据，开发人员和工程师可以对问题进行实时故障排除，提供更好的产品，并为预防措施提供警报系统。

物联网数据实时分析

互联网APP、物联网等实时应用程序监控、欺诈检测和实时排行榜等应用。

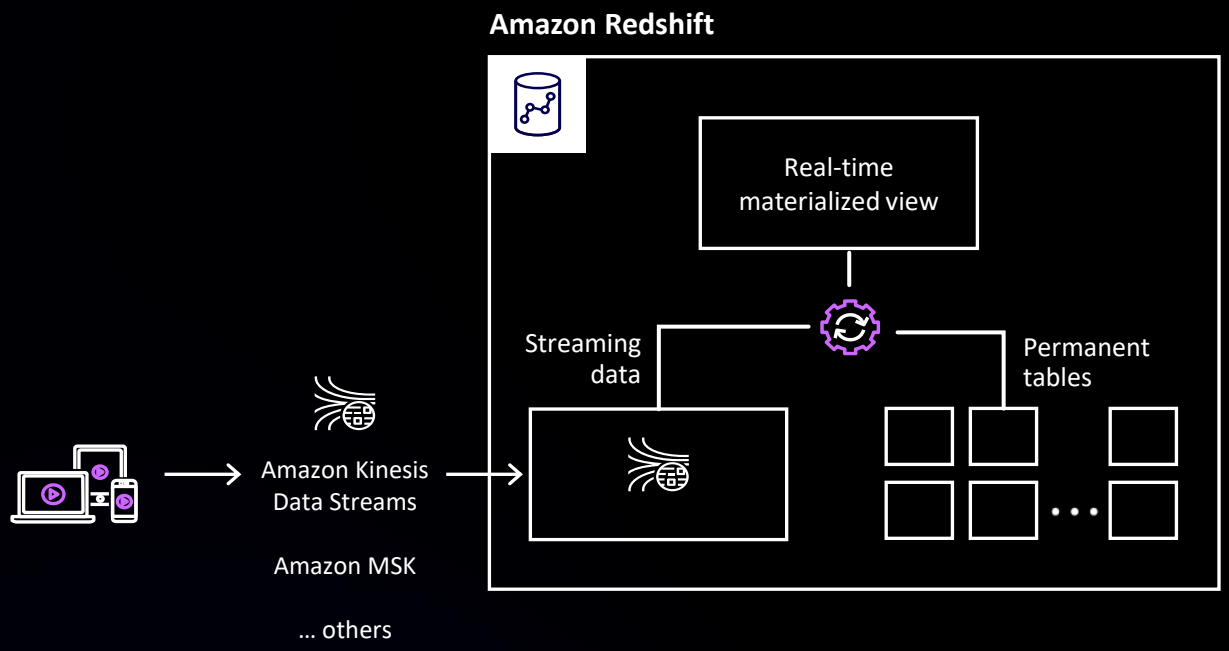
Redshift实时数据摄入Streaming Ingestion

与流引擎原生集成，实现快速流数据摄入

高速摄入
支持高达30万/秒的数据摄入（2KB size），小于30秒的延迟

易于使用
全SQL配置，直接将KDS中的数据实时摄入到Redshift

结构灵活
通过SUPER数据类型来摄入半结构化数据

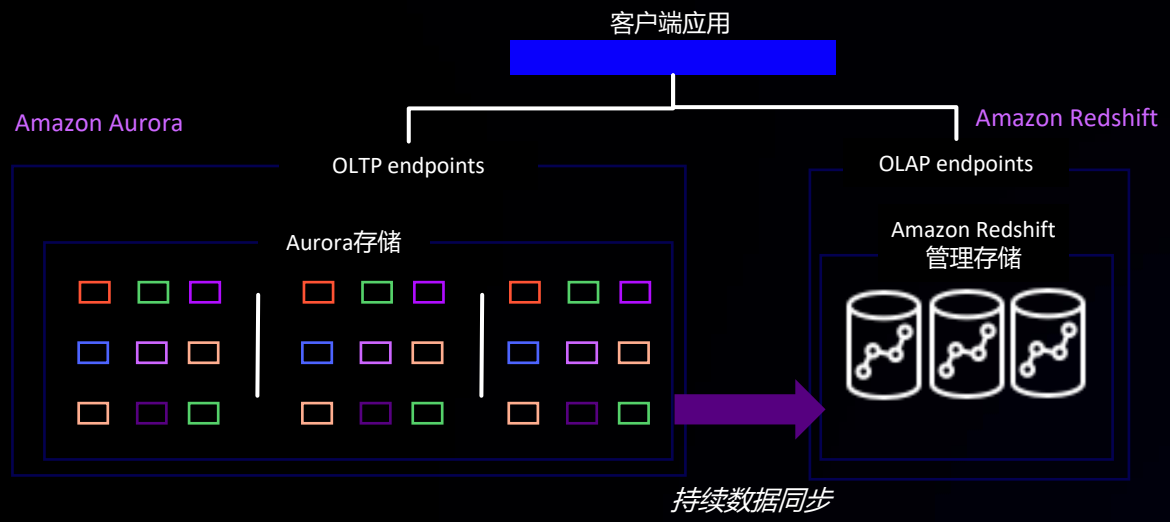


亚马逊云科技

新功能

新功能预览

Amazon Aurora与Redshift的zero-ETL数据实时同步



无需构建复杂的ETL任务流程

近实时的对Aurora中的交易数据进行分析与构建机器学习模型

零代码将多个Aurora数据库中的数据增量同步到一个统一的Redshift数据仓库中

亚马逊云科技

NEW

新功能发布

Amazon Redshift 与 Spark的集成

简化加速Spark任务直接访问Redshift中的数据

通过Java、 Python、 Scala来编写Apache Spark应用程序，来访问数据仓库中的数据

无需手动设置和维护开源Spark-Redshift connectors

通过仅将有关联的数据从Amazon Redshift移动到使用应用程序来提高任务性能

基于 IAM的凭证来提高安全性

亚马逊云科技

NEW

新功能预览

支持从S3自动加载数据

自动将S3上的文件加载到Redshift中

简单的低代码数据摄取

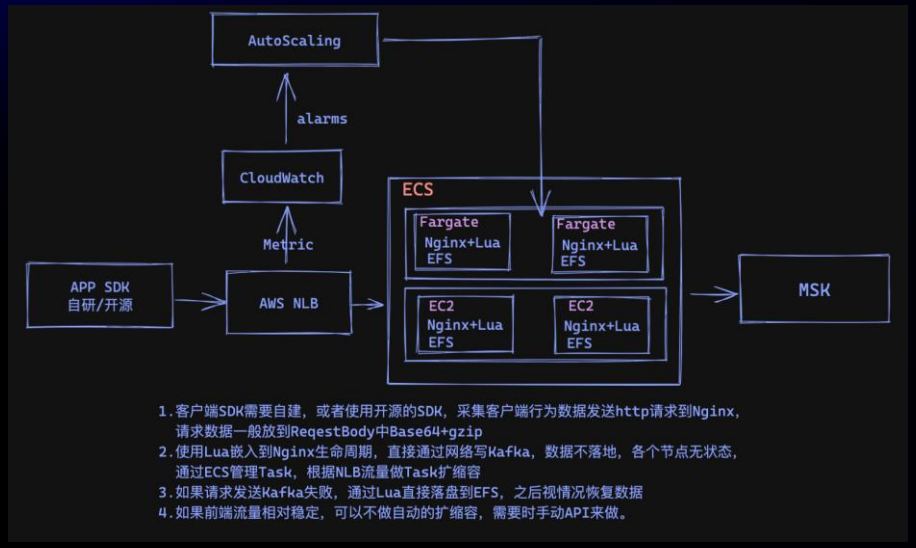
避免重复加载和手动跟踪加载的文件

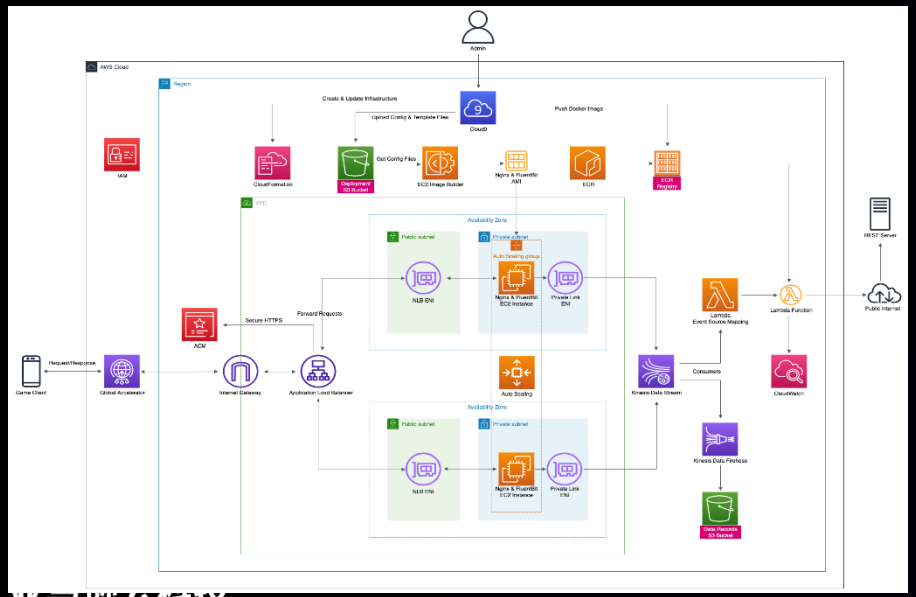
轻松将现有的 COPY 语句转换为自动摄取作业

用户可自定义配置从 Amazon S3 自动提取新数据

实时数仓参考架构 – APP埋点数据实时采集与分析

REDSHIFT STREAMING INGESTION与流引擎原生集成，实现快速流数据摄入



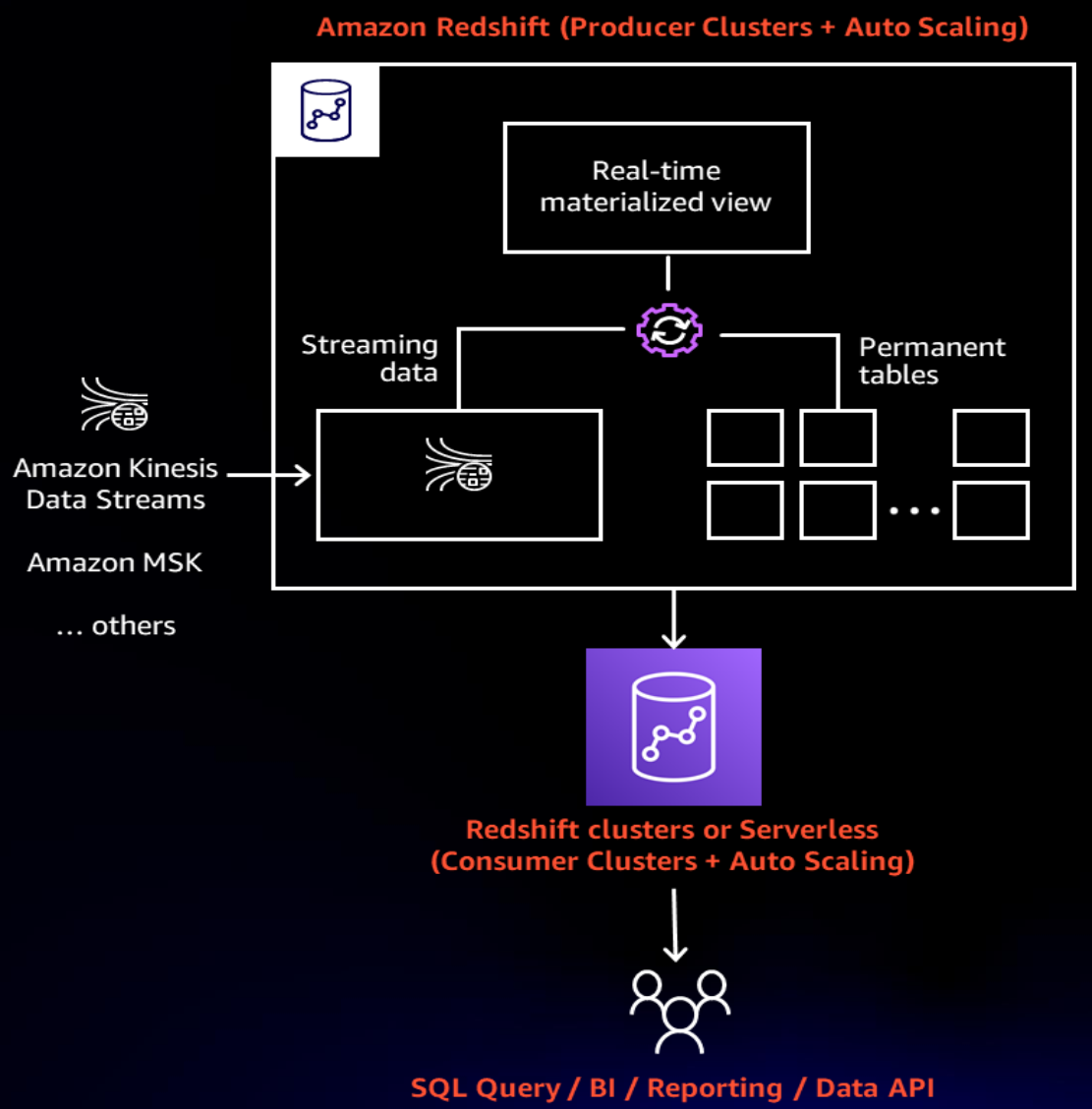


实时数据摄入
支持高达30万/秒的数据摄入（2KB size/row），小于10秒的延迟

高并发实时查询
支持大宽表、多表关联、复杂聚合等各种SQL查询，高并发，秒级响应

易于使用
全SQL配置，直接将KDS中的数据实时摄入到Redshift

结构灵活
通过SUPER数据类型来摄入半结构化数据



实时数仓参考架构 – 实时查询与实时计算

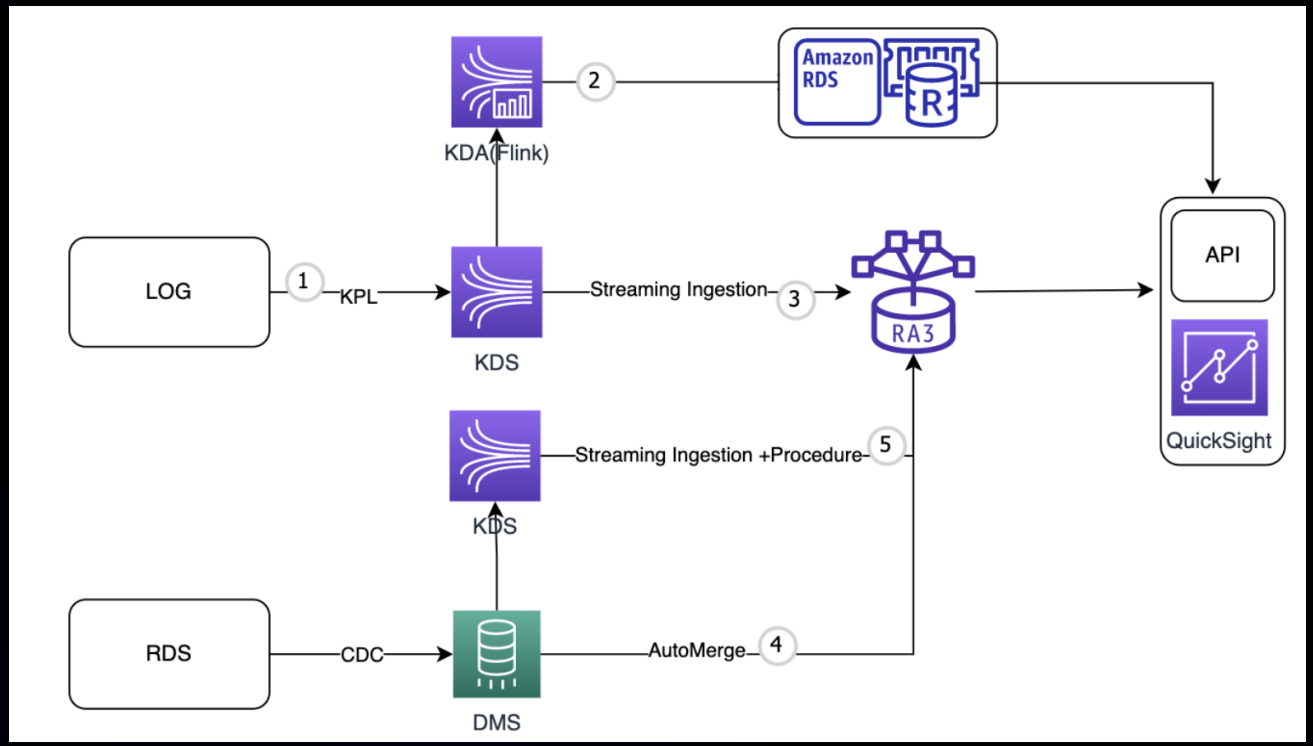
标号1
日志数据通过KPL或者Kinesis Agent发送到Kinesis Data Stream， （KDS是Serverless服务，支持API方式扩缩容）

标号2
端到端秒级延迟的数据，通过KDA(Flink Runtime)消费KDS中的数据，经过流计算后的结果Sink到RDS或者KV对外提供API查询。(KDA Flink Runtime是Serverless服务，KPU为计算单元，支持动态扩缩容)

标号3
通过Redshift Streaming Ingestion可以直接消费Kinesis(预计2022 Q2~Q3支持MSK)到Redshift，直接创建一个Kinesis的物化视图即可,秒级别数据延迟，30W/S吞吐(2kb size),小于30秒延迟。

标号4
对于RDS中的数据，DMS支持CDC同步，直接到Redshift中。比如MySQL开启Binlog，然后DMS同步Binlog信息到Redshift，支持Schema部分自动变更(比如源端增加列DMS+Redshift会自动同步变更)。 同时DMS也支持跑批全量同步数据

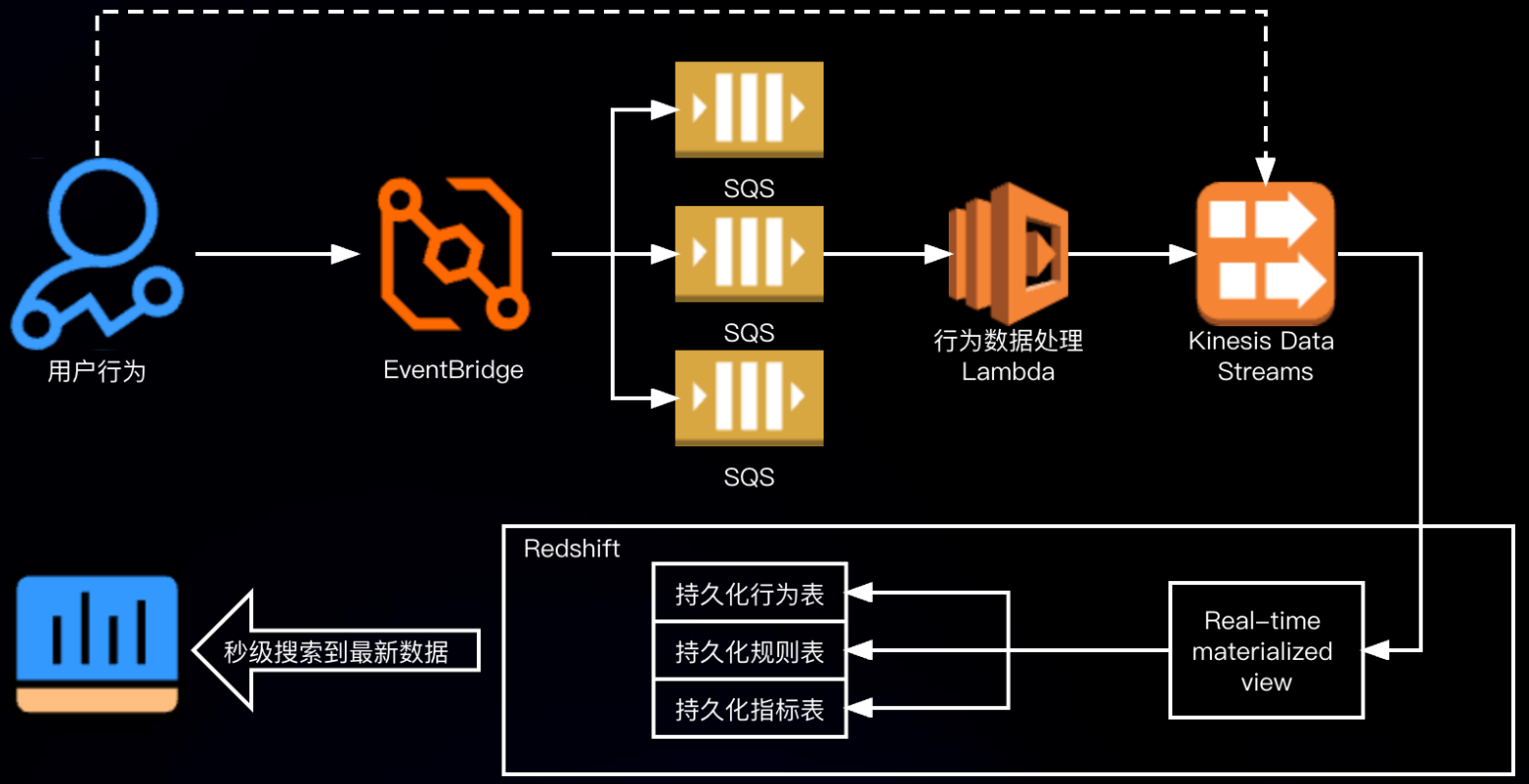
标号5
可以通过DMS或者其他CDC工具(比如Debezium， Flink CDC， Canal)将CDC数据发送到Kinesis，通过Redshift Streaming Ingestion+存储过程(Updata=Delete+Insert)来进行CDC数据实时写入Redsfhit。



亚马逊云科技

Redshift实时数仓应用场景 – 用户行为日志分析

- 存储从百亿扩充到千亿，成本降低50%
- 查询效率提高 30%，分钟级别搜索到满足条件的用户群体
- 数据分析效率显著提升
- 无服务轻松应对业务流量突增情况



今天开始测试 Amazon Redshift Serverless

$0.375
PER RPU-HOUR

在Redshift Serverless GA的同时，我们也向下调整了Redshift Serverless的价格，以US-EAST-1为例，每RPU小时的价格从$0.5美金调整为$0.375美金，价格降低25%!

$300
CREDIT

300美金的Credit来体验Redshift Serverless

亚马逊云科技

极致性能的数据仓库平台

Amazon Redshift

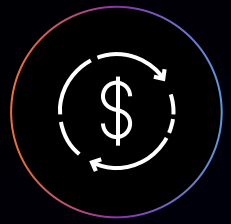
为云专门构建

简单易用的分析

快速洞察，无需关心基础架构

分析所有数据

与数据湖深入集成



任意规模性能

提供超过3x的性价比

亚马逊云科技


openGauss
OceanBase
ArkDB
RASESQL
HotDB
StellarDB
QianBase xTP
云树Shard
GoldenDB
DolphinDB
MatrixDB
DynamoDB
SinoDB
FastData
Galaxybase
KunDB
SQL Server
vertica
D B 2
GBase
Oracle
达梦数据库
Spacture
开务数据库
UbiSQL
神舟通用
TDSQL
TiDB
Tapdata
StarRocks
KingbaseES
2010
2014
2018

THANKS