

云数据库HybridDB for PostgreSQL -- 混合云海量互联网广告分析

2017年01月05日 萧少聪(铁庵)

[HybridDB(基于Greenplum Database深度优化)]

如何实现100倍互联网业务分析性能提升

打造OLAP混合云数据仓库方案

阿里云Postgres生态体系

HybridDB(基于Greenplum Database深度优化)

Pivotal

IT大咖说
知识分享平台

2008年：Greenplum商业版本正式进入中国，主要面向重度分析需求的企业，如：

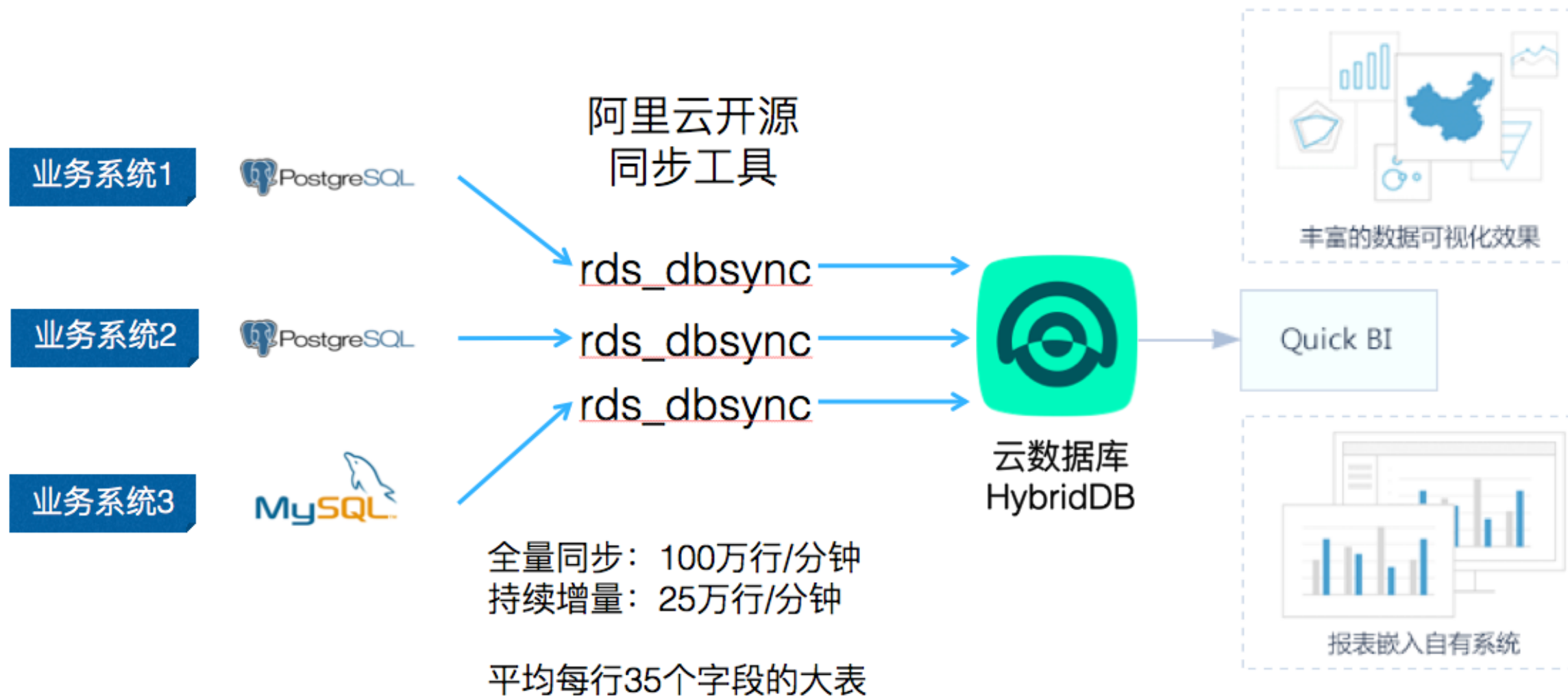
- 金融业、电信、航空、物流、制造业等行业

2016年07月：阿里云Greenplum版本正式公测

2016年12月：商业化发布重命名为 云数据库HybridDB，在GPDB基础上扩展支持

- 插件式外部组件管理
- JSON数据类型 / 优化GIS处理（用于互联网或物联网分析）
- HyperLogLog预估分析数据类型（多用于联系网广告分析的PV、UV预估）
- 基于OSS云存储的外部表（一份数据打通 OLTP / OLAP / BigData 业务）
- 阿里云开源 rds_dbsync 数据增量同步工具

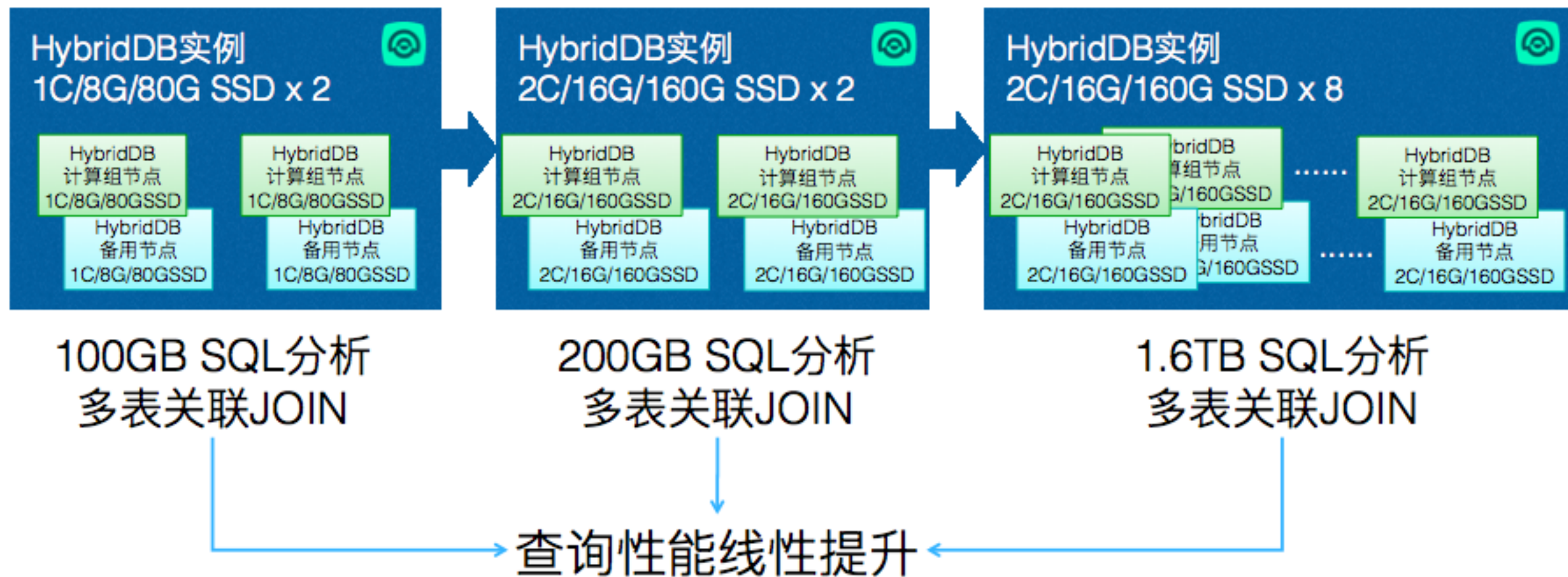
HybridDB(基于Greenplum Database深度优化)



HybridDB(基于Greenplum Database深度优化)

Pivotal

IT大咖说
知识分享平台





基本信息

帐号管理

数据库连接

监控与报警

数据安全性

基本信息

[设置白名单](#)

实例ID: [REDACTED]

名称: [REDACTED]

地域可用区: 华北 2可用区C

实例类型: 专享实例

内网地址: [REDACTED] [复制地址](#)

内网端口: 3432

外网地址: [REDACTED] [复制地址](#)

外网端口: 3432

运行状态

[释放](#)

运行状态: 运行中

付费类型: 按量付费

创建时间: 2017-02-07 14:46:02

配置信息

[变更配置](#)

数据库类型: HybridDB 4.3

计算组规格: gpdb.group.segsdx2

计算组详情: 2Core(s) / 16384 MB 内存 /160 GB SSD / 300 MBPS

计算组数量: 2

计算资源汇总: 4Core(s) / 32768 MB 内存 /320 GB SSD / 600 MBPS

基本信息

帐号管理

数据库连接

监控与报警

数据安全

监控汇总

计算组监控

资源监控:

资源监控

刷新

查询时间:

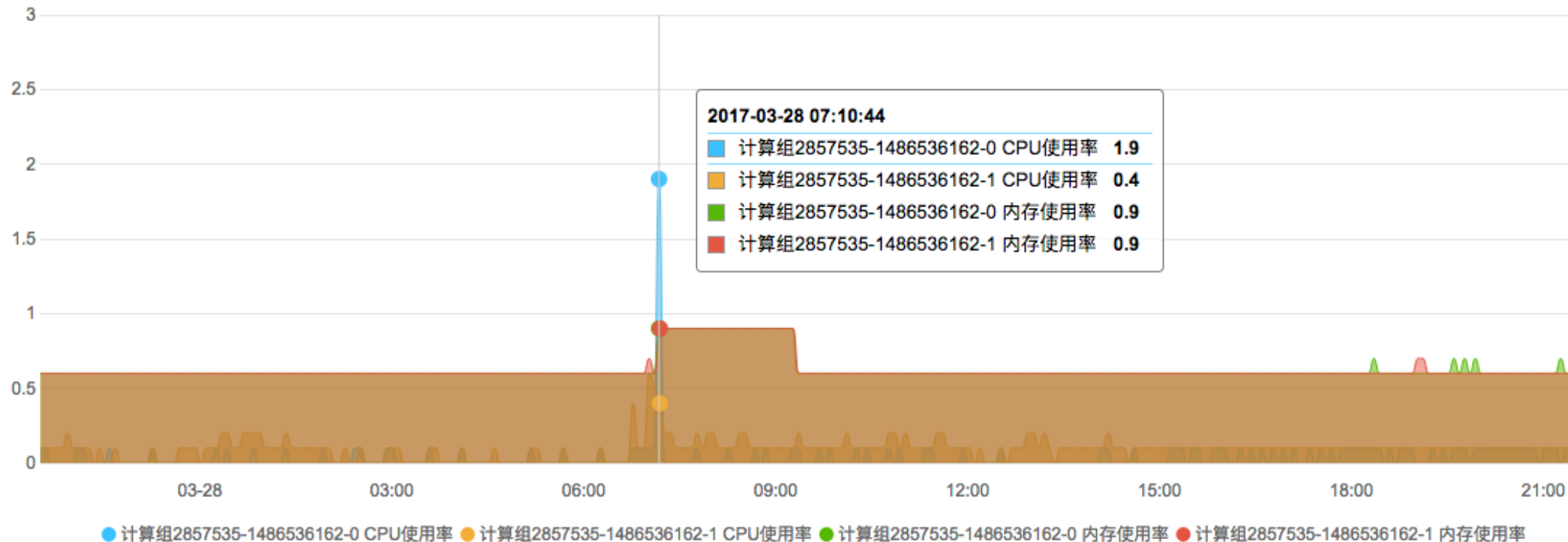
1小时

1天

2017-03-27 21:29 - 2017-03-28 21:29

≡

CPU和内存利用率 (%)



读IO吞吐量 单位(MB/S)

针对不同用户的需求 阿里云 支持了丰富的 Greenplum 插件

- PostGIS: 支持地理信息数据。
- MADlib: 机器学习方面的函数库。
- fuzzystmatch: 字符串模糊匹配。
- orafunc: 兼容 Oracle 的部分函数。
- oss_ext: 支持从 OSS 读取数据。
- hll: 支持用 HyperLogLog 算法进行统计。
- pljava: 支持使用 PL/Java 语言编写用户自定义函数 (UDF) 。
- pgcrypto: 支持加密函数。
- intarray: 整数数组相关的函数、操作符和索引支持

HybridDB (基于Greenplum Database深度优化)

[如何实现100倍互联网业务分析性能提升]

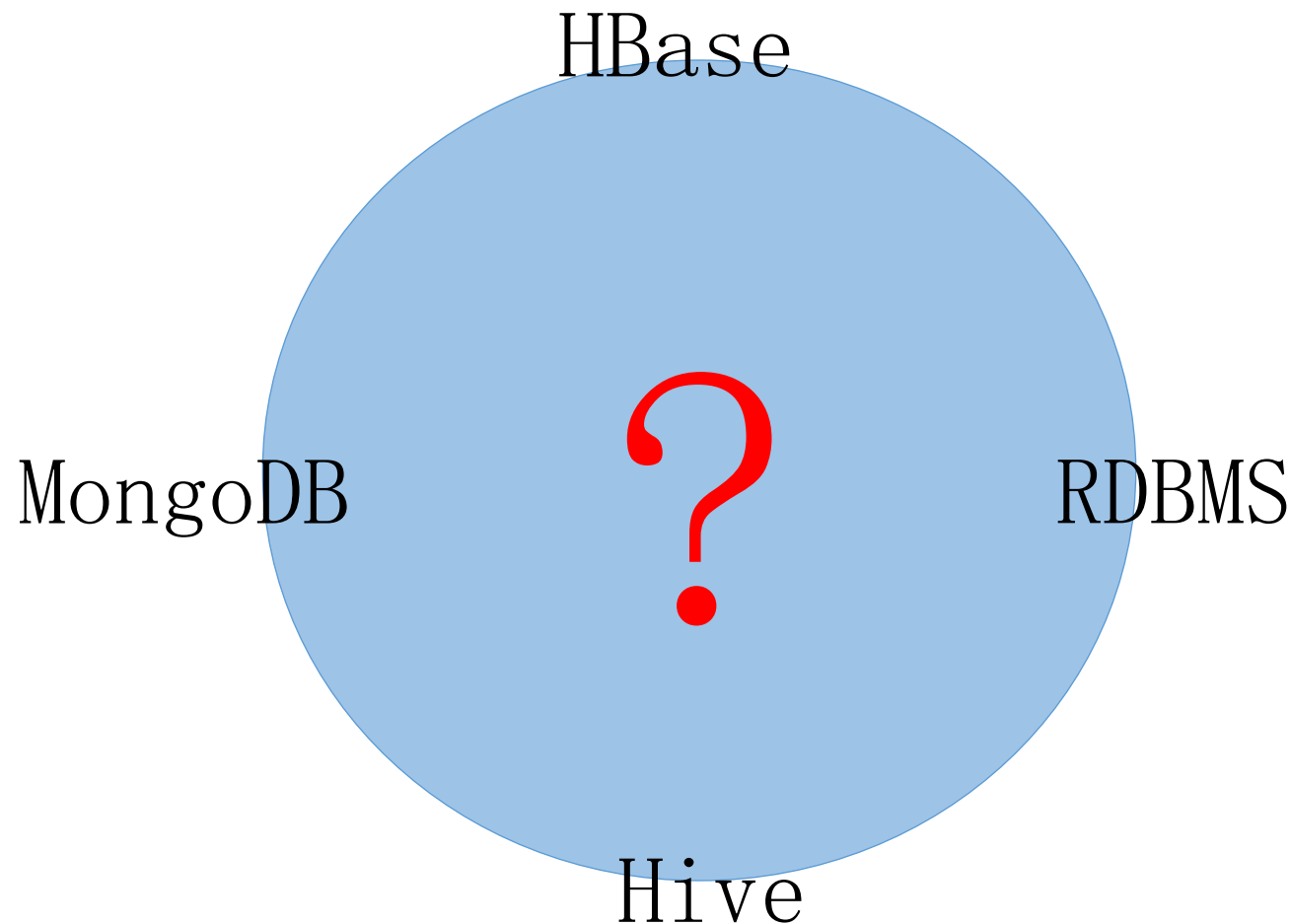
打造OLAP混合云数据仓库方案

阿里云Postgres生态体系

HybridDB for PostgreSQL 支持JSON助力互联网&物联网

HybridDB

IT大咖说
知识分享平台



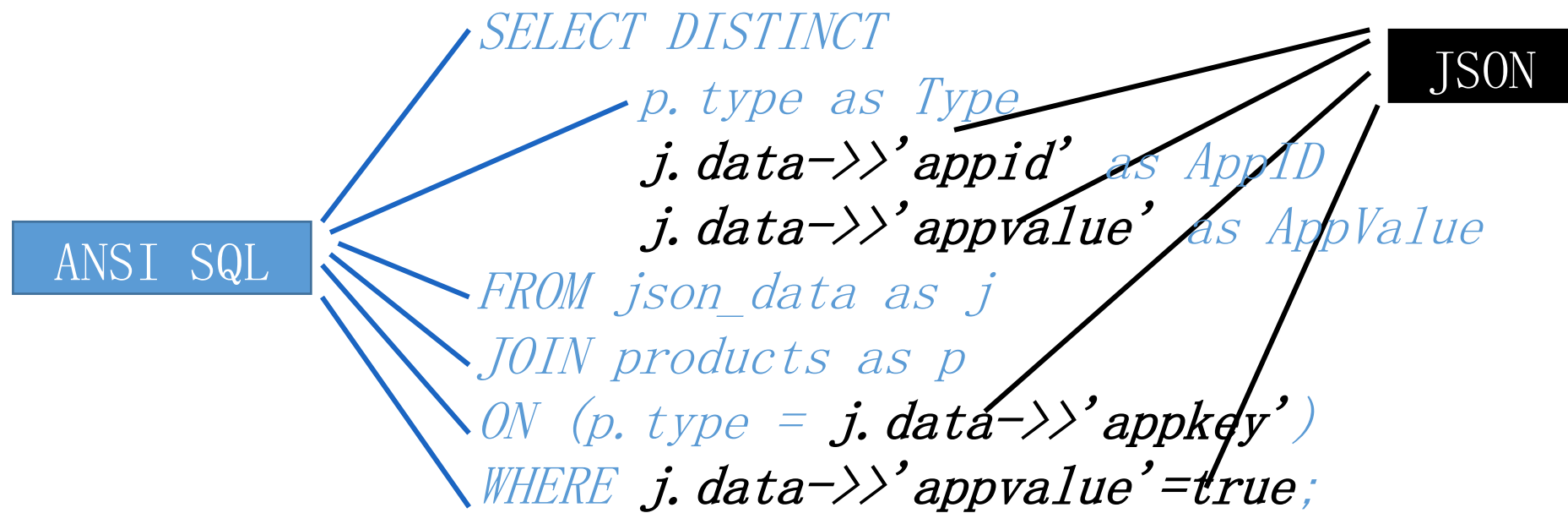
HybridDB for PostgreSQL 支持JSON助力互联网&物联网

HybridDB

IT大咖说
知识分享平台

```
CREATE TABLE json_data(data json);
```

```
INSERT INTO json_data VALUES ('{"appid":"1234","appkey":"xx","appvalue":"true"}');
```



HybridDB for PostgreSQL 支持IPv4/v6网络类型为广告分析提速



Name	Storage Size	Description
cidr	12 or 24 bytes	IPv4 and IPv6 networks
inet	12 or 24 bytes	IPv4 and IPv6 hosts and networks
macaddr	6 bytes	MAC addresses

- 专用的数据类型
- 专用的索引
- 专用的处理函数

Function	Return Type	Description	Example	Result
abbrev(inet)	text	abbreviated display format as text	abbrev(inet '10.1.0.0/16')	10.1.0.0/16
abbrev(cidr)	text	abbreviated display format as text	abbrev(cidr '10.1.0.0/16')	10.1/16
broadcast(inet)	inet	broadcast address for network	broadcast('192.168.1.5/24')	192.168.1.255/24
family(inet)	int	extract family of address; 4 for IPv4, 6 for IPv6	family('::1')	6
host(inet)	text	extract IP address as text	host('192.168.1.5/24')	192.168.1.5
hostmask(inet)	inet	construct host mask for network	hostmask('192.168.23.20/30')	0.0.0.3
masklen(inet)	int	extract netmask length	masklen('192.168.1.5/24')	24
netmask(inet)	inet	construct netmask for network	netmask('192.168.1.5/24')	255.255.255.0
network(inet)	cidr	extract network part of address	network('192.168.1.5/24')	192.168.1.0/24
set_masklen(inet, int)	inet	set netmask length for inet value	set_masklen('192.168.1.5/24', 16)	192.168.1.5/16
set_masklen(cidr, int)	cidr	set netmask length for cidr value	set_masklen('192.168.1.0/24'::cidr, 16)	192.168.0.0/16
text(inet)	text	extract IP address and netmask length as text	text(inet '192.168.1.5')	192.168.1.5/32

在互联网中，经常需要计算UV和PV。所谓PV即Page View，网页被打开多少次，如视频网站非常重视视频的点击率，即被播放多少次，也即PV。而UV即Unique Visitor如朋友圈或者微信公众号中的文章则统计有多少人看过该文章，也即UV。

? count distinct 还是 count group by ?

上亿的数据量根本无法进行此操作

PV、UV是否需要100%准确

HybridDB for PostgreSQL 广告行业高效PV、UV估算

```
=> CREATE EXTENSION hll;
```

```
=> CREATE TABLE access_date_hll (acc_date date unique, userids hll);
```

----- 写入1亿数据 -----

```
=> select #userids from access_date_hll where acc_date='2017-02-03';  
      ?column?
```

```
-----  
1515803.229769629  
(1 row)
```

Time: 36.519 ms

误差: 1%

```
=> create table access_date_bigint (acc_date date, userids bigint);  
=> select count(distinct userids) from access_date_bigint where  
acc_date=current_date-1;  
      count
```

```
-----  
1500001  
(1 row)
```

Time: 96324.984 ms

HybridDB for PostgreSQL 100倍性能提升不是梦

Pivotal

IT大咖说
知识分享平台

```
select count(distinct
```

```
      : 44 min 35.96 s  
gp: 8.6s
```

相差311倍

```
select sum(money*all;
```

```
      : 33s  
gp: 2.5s
```

相差13倍

```
select sum(money) as
```

```
      : 72.7s  
gp: 1.7s
```

相差43倍

```
select sum(case when  
sum(case when type=1  
sum(case when type=2  
sum(case when type=3  
sum(case when type=0  
sum(case when type=1  
sum(case when type=2  
sum(case when type=3  
from lzh_borrowlist v
```

```
      : 35s  
gp: 55s
```

```
set optimizer = on;之
```

```
gp: 6s
```

相差6倍

HybridDB (基于Greenplum Database深度优化)

如何实现100倍互联网业务分析性能提升

[打造OLAP混合云数据仓库方案]

阿里云Postgres生态体系

统一接口打通OLTP/AP，云上云下通用

Pivotal

IT大咖说
知识分享平台



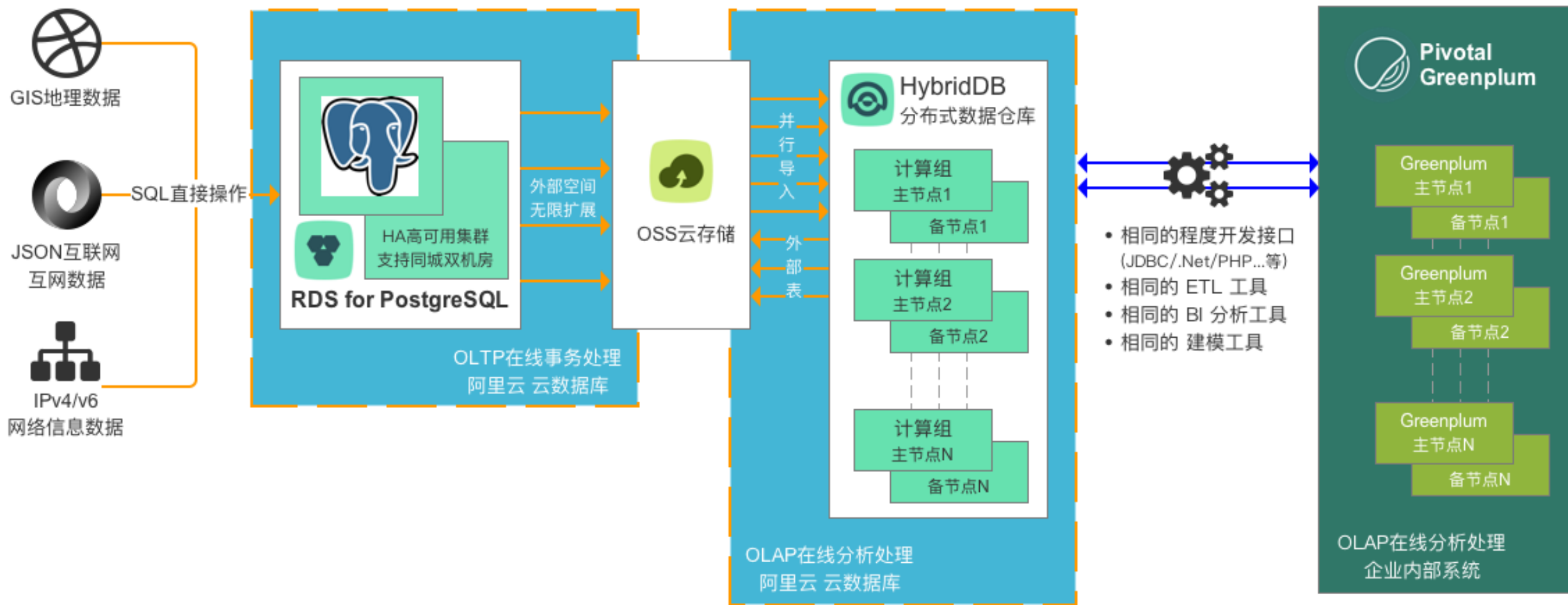
用户IDC机房



统一接口打通OLTP/AP, 云上云下通用

Pivotal

IT大咖说
知识分享平台



HybridDB (基于Greenplum Database深度优化)

如何实现100倍互联网业务分析性能提升

打造OLAP混合云数据仓库方案

[阿里云Postgres生态体系]

覆盖70%以上关键数据库在线引擎

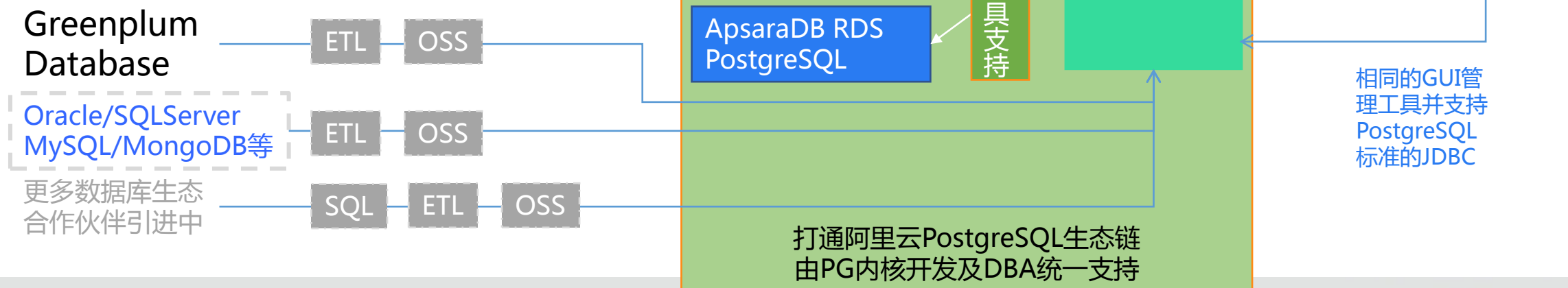


Postgres不是一个数据库，Postgres是一整套生态

Pivotal

IT大咖说
知识分享平台

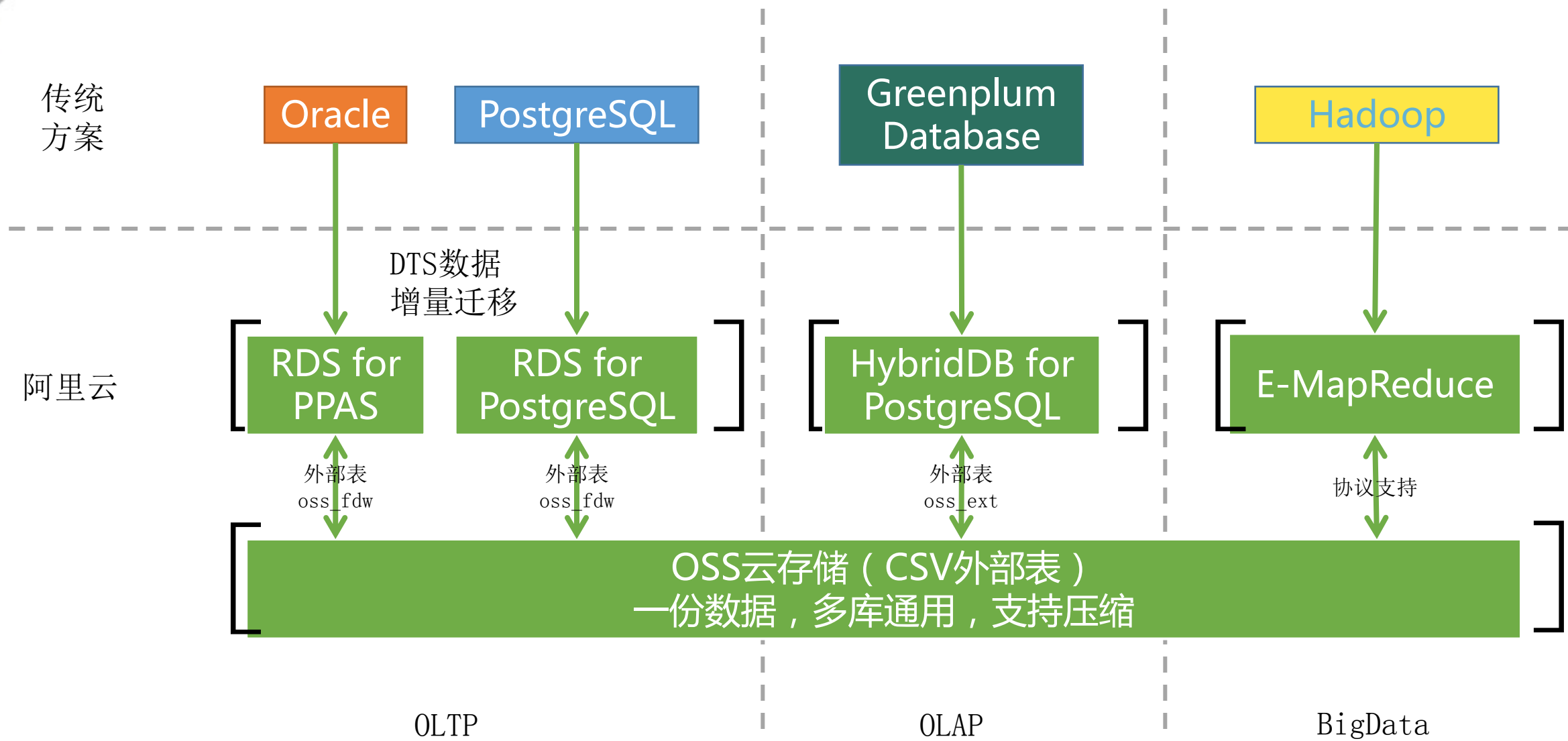
- **SQL操作流**：易于用户迁移
SQL92/99/08集合Oracle兼容
- **数据流**：众多ETL工具支持，未来还会打通DTS等增量渠道
- **国际化对接**：使用与AWS数据仓库Redshift相同的接口驱动



阿里云 基于PostgreSQL从OLTP->OLAP->BigData

Pivotal

IT大咖说
知识分享平台



1. PostgreSQL操作OSS

https://help.aliyun.com/document_detail/44461.html

2. PPAS操作OSS同上，但进行CREATE EXTENSION时参考

https://help.aliyun.com/knowledge_detail/42699.html

3. HybridDB操作OSS:

https://help.aliyun.com/document_detail/35457.html

4. EMR操作OSS:

https://help.aliyun.com/document_detail/42799.html

HybridDB (基于Greenplum Database深度优化)

如何实现100倍互联网业务分析性能提升

打造OLAP混合云数据仓库方案

阿里云Postgres生态体系

为了无法计算的价值 | 阿里云

