MaxCompute基于BigBench标准的最新测试进展

阿里云技术专家　路璐

敢为人先、引领潮流 BigBench On MaxCompute 2.0

100 TB

首个100TB测试通过的引擎
The first engine passed 100TB Bigbench verification

7830 QPM

首个达到7000分的引擎
The first engine reaches 7000+

$371.9/QPM 预付费包3年价格 Pre-paid for 3yrs

$12.3/QPM 预付费包1月价格 Pre-paid for 1 month

$2.1/QPM 按需后付费价格 Post paid by usage

首个基于公共云服务的Benchmark
The first BigBench verification run on public cloud

Disclaimer:
the BigBench kit used by these performance tests is derived from TPCx-BigBench.

TPCx-BB简介

业界领先的基于端到端的大数据分析领域应用级测试基准

由英特尔领衔发起、主要开发和大力推广

大数据工业应用特征：SLA，TCO

多种任务类型：SQL，MapReduce，MachineLearning，Streaming

完整的软硬件性能评估标准：Scale，BBQpm， Price/BBQpm

10TB

当前最大数据规模

1491.23 BBQpm

当前最佳性能

$589 Price/BBQpm

当前最高性价比

当前最佳成绩

TPCx-BB作业特征分布

Data Sources	Number of Queries
Structured	18
Semi-Structured	7
Un-Structured	5

Analytic techniques	Number of Queries
Statics analysis	6
Data mining	17
Reporting	8

Query Types	Number of Queries
Pure sql	13
Machine learning	5
OpenNLP	5
Java MR	3
Streaming	4

Query	Input DataType	Processing Model	Query	Input DataType	Processing Model
#1	Structured	Java MR	#16	Structured	Pure SQL
#2	Semi-Strunctred	Java MR	#17	Structured	Pure SQL
#3	Semi-Strunctred	Streaming	#18	UnStructured	OpenNLP
#4	Semi-Strunctred	Streaming	#19	UnStructured	OpenNLP
#5	Semi-Strunctred	Machine Learining	#20	Structured	Machine Learining
#6	Structured	Pure SQL	#21	Structured	Pure SQL
#7	Structured	Pure SQL	#22	Structured	Pure SQL
#8	Semi-Strunctred	Pure SQL	#23	Structured	Pure SQL
#9	Structured	Pure SQL	#24	Structured	Pure SQL
#10	UnStructured	OpenNLP	#25	Structured	Machine Learining
#11	UnStructured	OpenNLP	#26	Structured	Machine Learining
#12	Semi-Strunctred	Pure SQL	#27	UnStructured	OpenNLP
#13	Structured	Pure SQL	#28	UnStructured	Machine Learining
#14	Structured	Pure SQL	#29	Structured	Streaming
#15	Structured	Java MR	#30	Semi-Strunctred	Streaming

TPCx-BB性能评估标准

根据软硬件性能评估（SLA）

通过BBQpm@SF评估性能

涵盖数据规模、load测试阶段、Power测试阶段、Throughput测试阶段

$$BBQpm@SF = \frac{SF * 60 * M}{T_{ld} + \sqrt[2]{T_{pt} * T_{tt}}}$$

Load测试阶段： 测试数据导入

$$T_{ld} = 0.1 * T_{load}$$

Power测试阶段：单个查询流顺序执行30个查询语句

$$T_{pt} = M * \sqrt[M]{\prod_{i=1}^{i=M} Q(i)}$$

Throughput测试阶段：多个并发查询流并发执行查询语句

$$T_{tt} = \frac{1}{n} * T_{Tput}$$

M：查询流数量(30个)

SF (Scale Factor): 表示基准测试数据规模大小,比如1000代表1T

根据软硬件性价比评估（TCO）

通过$/BBQpm@SF评估性价比

$$\$/BBQpm@SF = \frac{C}{BBQpm@SF}$$

C：被评估SUT的总价格

BigBench on MaxCompute软件栈

BigBench on MaxCompute基于BigBench进行修改，兼容所有语义

软件栈完整，覆盖BigBench所有功能点

Query Type	Hive on Spark	MaxCompute
Pure sql	Hive Sql	MaxCompute SQL
Machine learning	Spark MLlib	PAI
OpenNLP	SQL+UDF	SQL + UDF
Java MR	SQL+UDF	SQL + UDF
Streaming	Python streaming	SQL + UDF

MaxCompute Hive兼容模式： 完全兼容开源hive的所有数据类型和SQL语法

Tunnel：MaxCompute数据导入导出工具，可以将不同平台数据导入到MaxCompute中

PAI：阿里巴巴一站式的机器学习平台，与MaxCompute数据打通

BigBench on MaxCompute结果分析

MaxCompute海量数据处理能力，总数据量达到EB级

基于阿里云自主研发的Apsara分布式操作系统，单集群机器规模达到万台

Fuxi：分布式资源管理和调度系统，2015年GraySort纪录

Pangu：大规模分布式文件系统，可以支持10亿+文件

MaxCompute新一代执行引擎，从Compiler、Optimizer、Runtime等模块进行深度优化

Range partition、AutoMapjoin、ShuffleRemove等优化方法

与Intel全面合作，软硬结合深度优化， 充分发挥至强®可扩展处理器架构优势

第一个公共云服务Benchmark

完整的TCO，包含软硬件和运维服务

计价规则灵活，可以按需后付费、按月预付费、按年预付费等

规模和性能优势

数据规模达到100TB

性能达到7830Qpm

性价比达到$2.1/Qpm

开放一个月测试期，免费提供MaxCompute测试资源

开源Test Kit并提供操作指南

用户可以在MaxCompute平台上测试验证BigBench

Contact us:

email：lu.lu@alibaba-inc.com

钉钉扫码

THANK YOU

Scan QR Code

关注MaxCompute产品社区
Community

了解MaxCompute产品详情
Product Details

加入MaxCompute钉群咨询
Join DingTalk Group

诚聘MaxCompute英才
We are hiring!