

MaxCompute性能优化实践

阿里巴巴集团-计算平台事业部 路璐

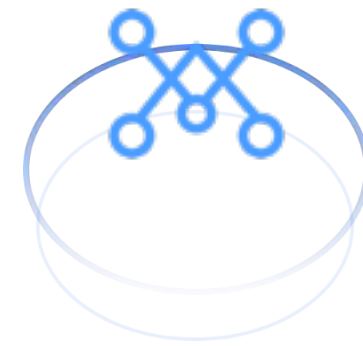
MaxCompute/ODPS -- 阿里巴巴和阿里云大数据的旗舰计算平台



99%存储 + 95%计算
阿里巴巴内部统一的
大数据平台，支持阿里所有业务



BigBench 2.5X
高性能，低成本



60K+/10+
超大规模
跨DC调度容灾能力



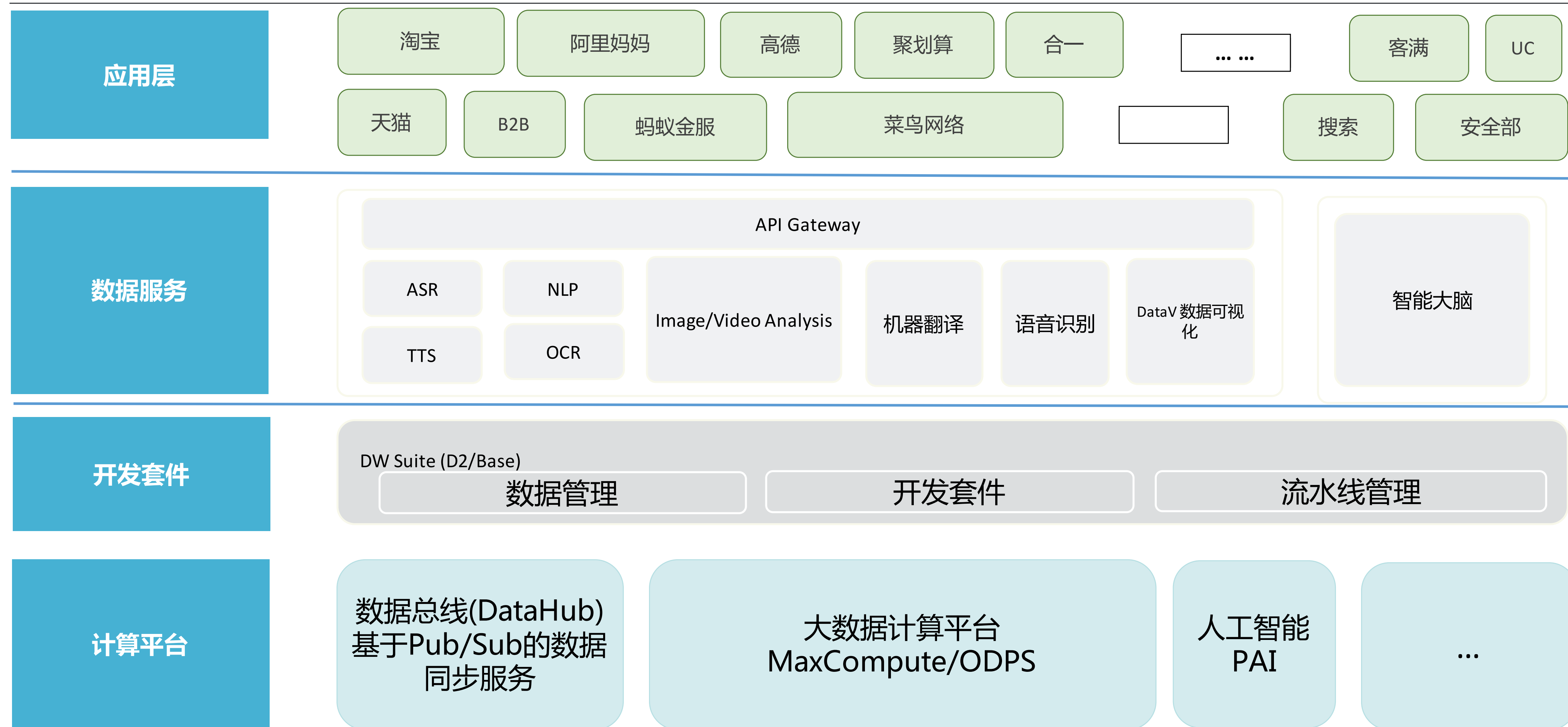
250% X
大数据旗舰平台
公共云支撑上层“大脑”和数加



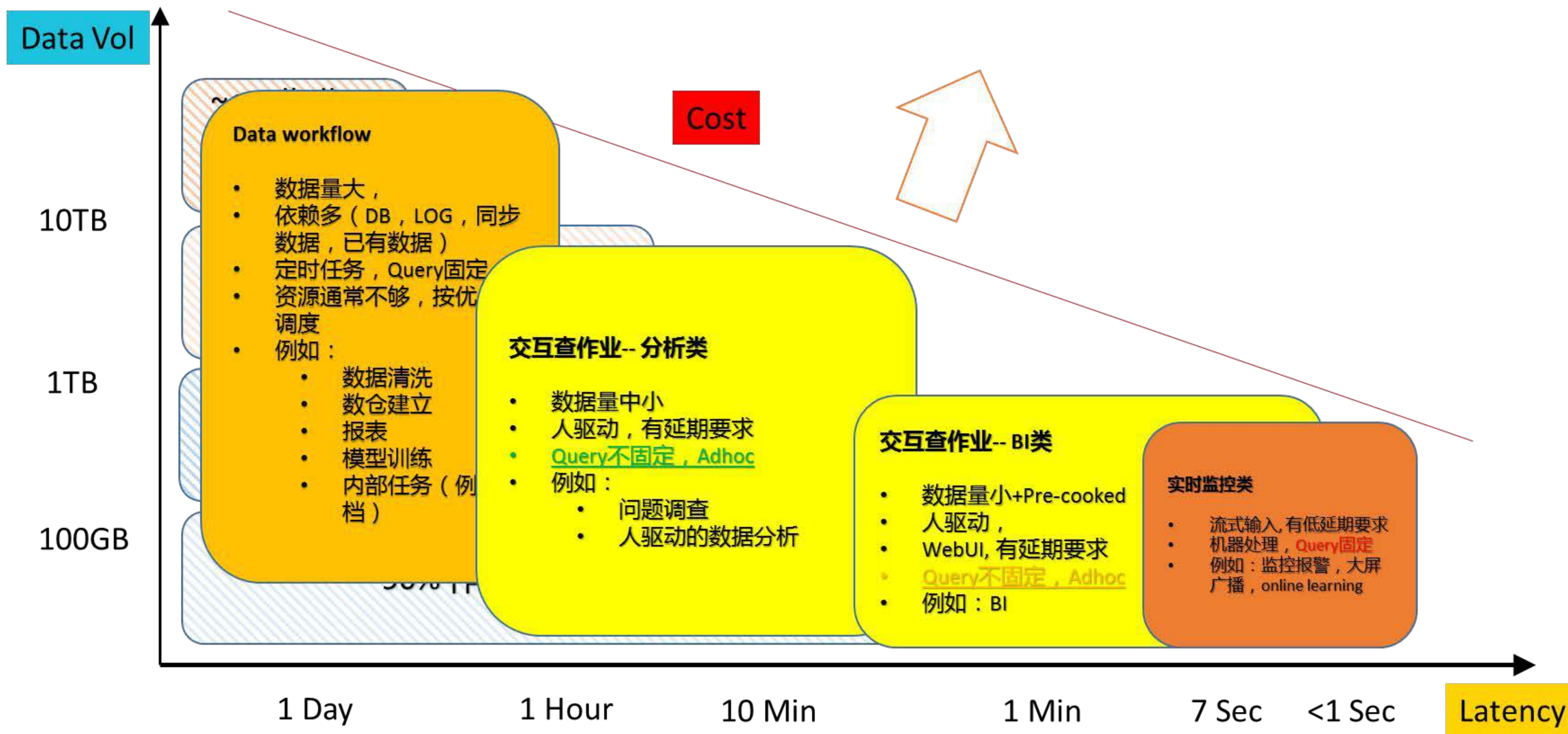
50套+
作为大数据旗舰平台
专有云部署到各行各业



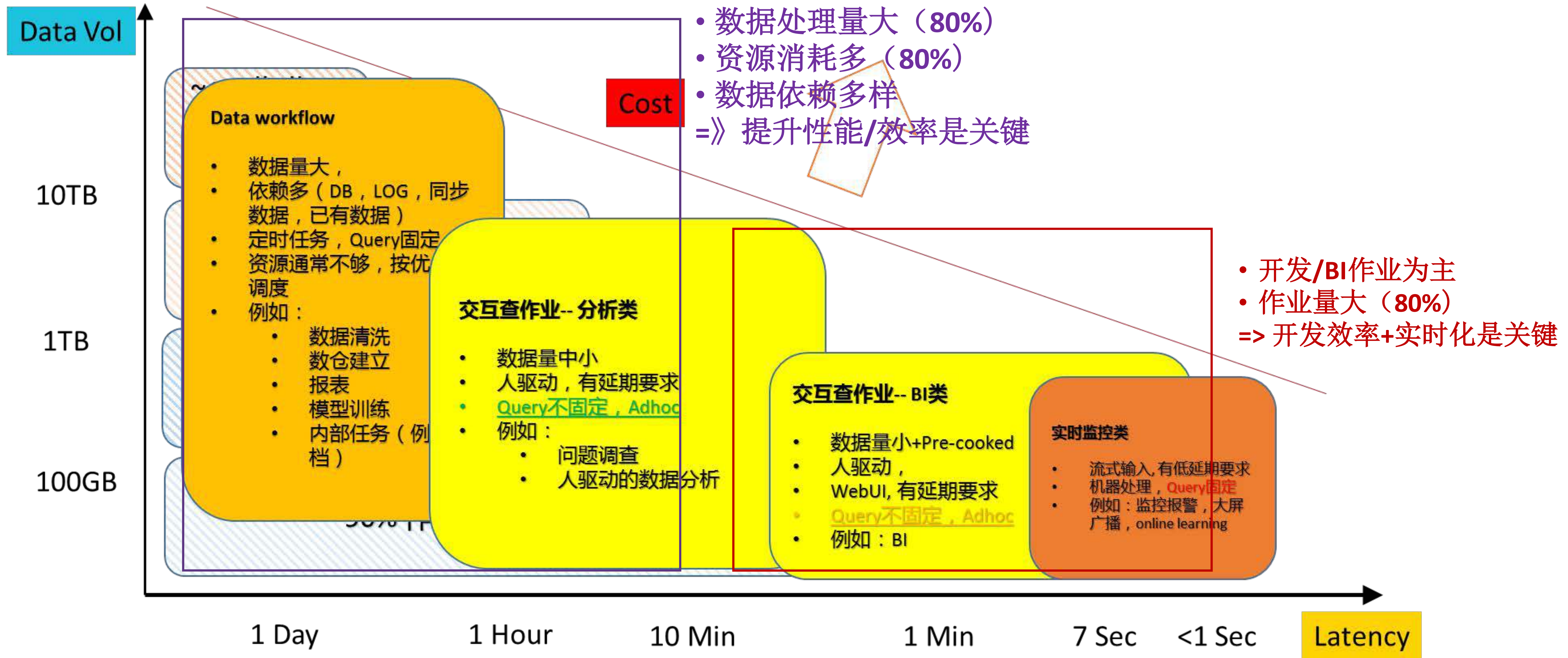
阿里云大数据计算服务 (MaxCompute/ODPS)



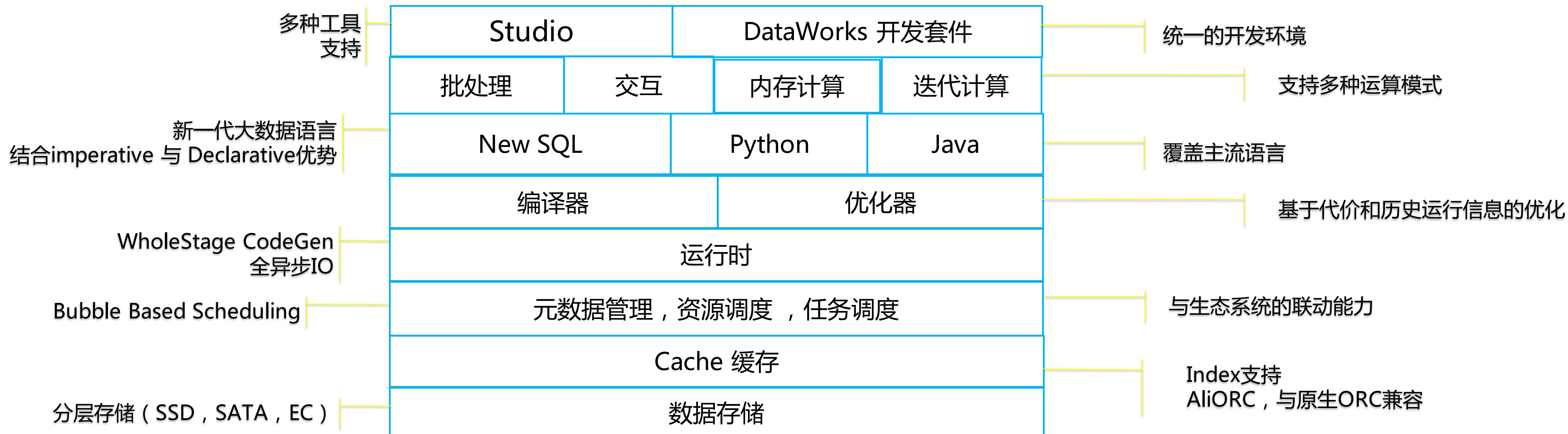
大数据计算 典型场景分析 (从计算量和延迟的角度)



大数据计算 典型场景分析 (从计算量和延迟的角度)



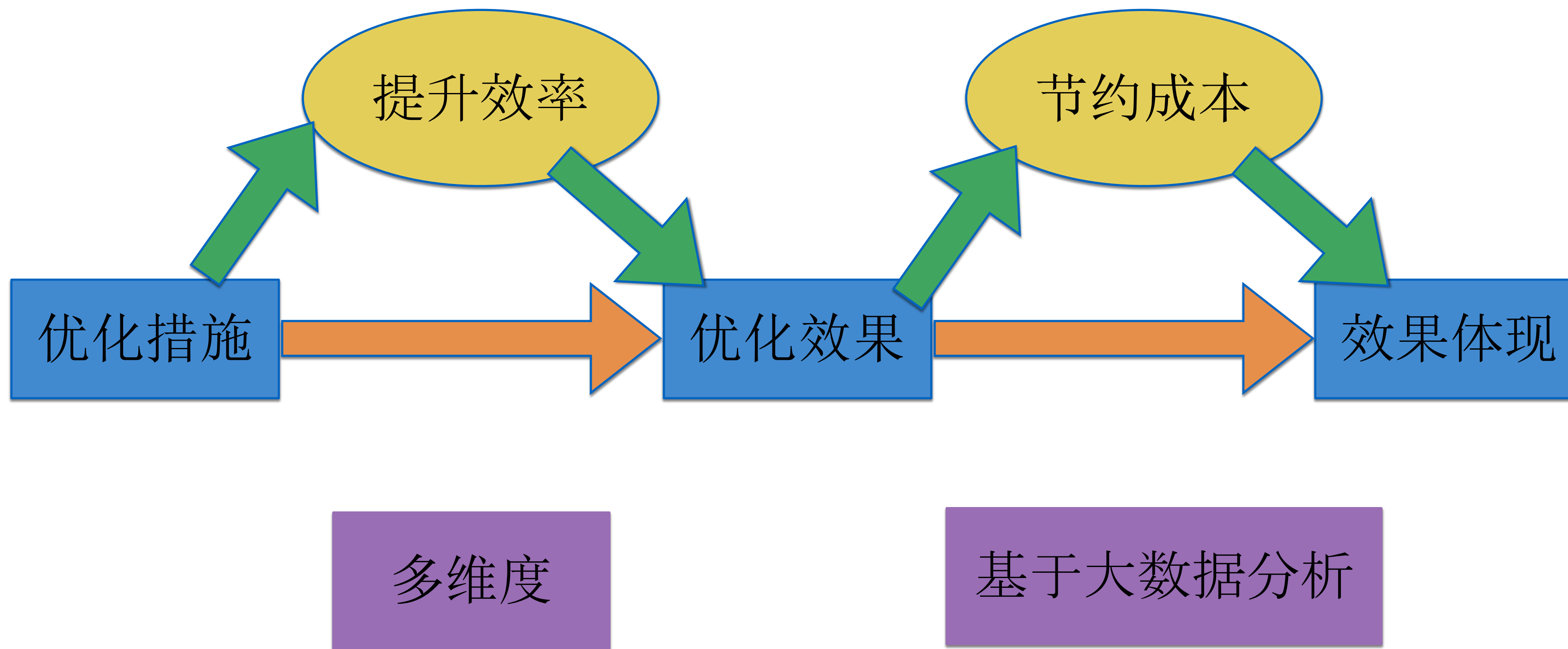
MaxCompute 2.0 架构持续升级



MaxCompute 2.0 架构优化——HBO

HBO (History-Based Optimization) 是基于任务执行历史的优化方式。

任务执行历史 + 集群状态信息 + 优化规则 -> 更优的执行配置

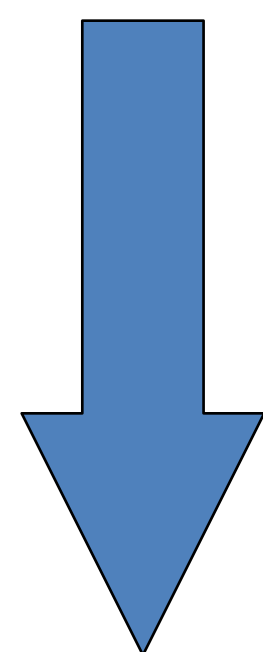


MaxCompute 2.0 架构优化——runtime行转列



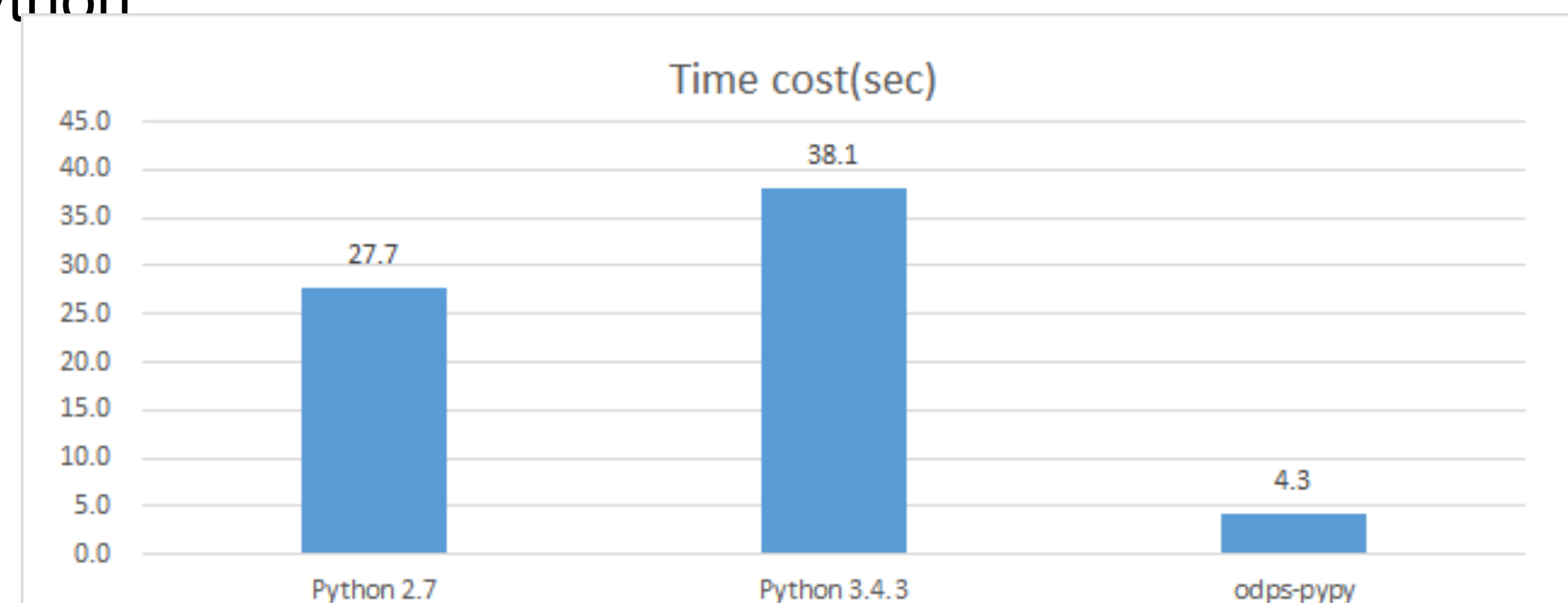
MaxCompute 2.0 架构优化——python udf

Python 占比太高

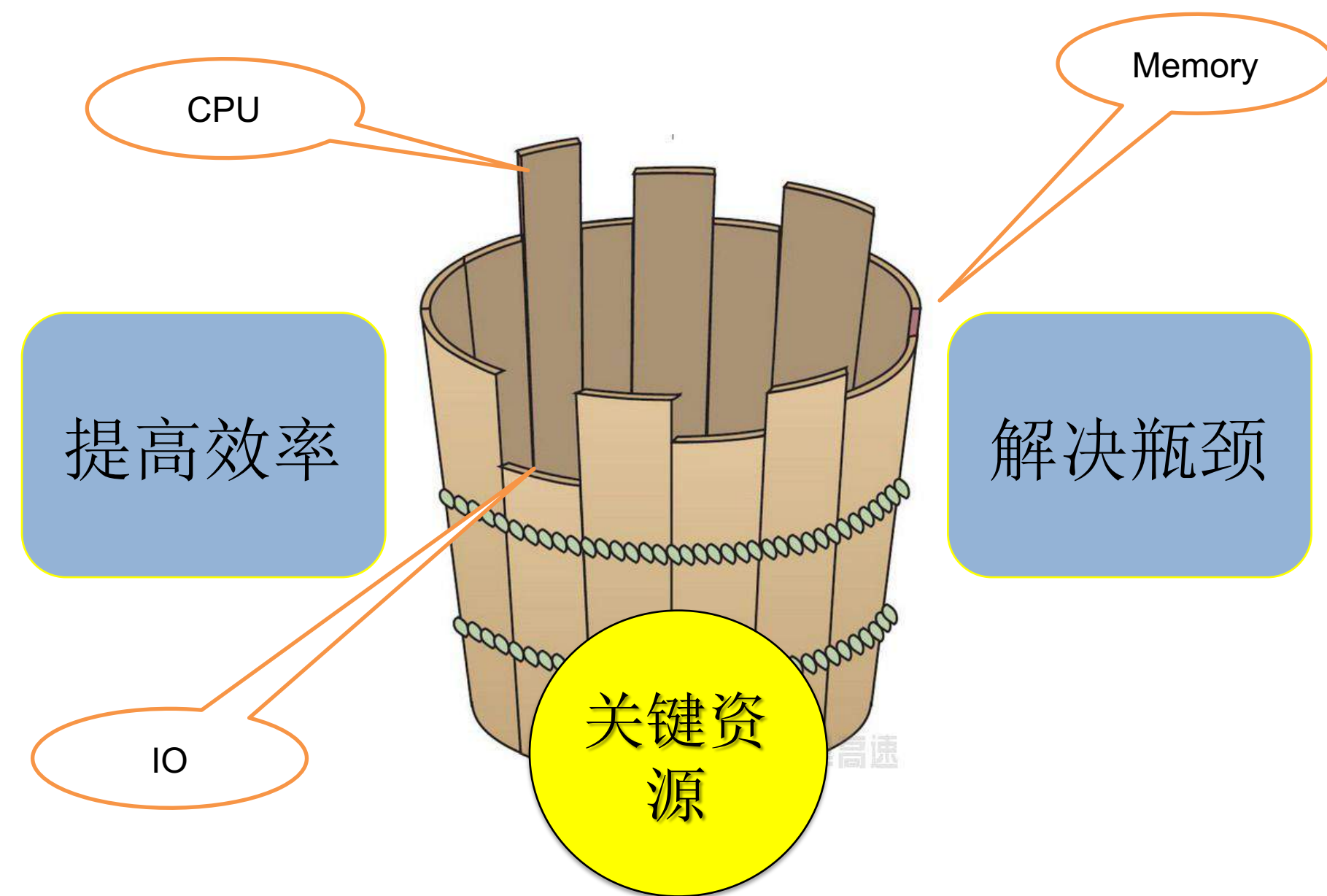


PyPy

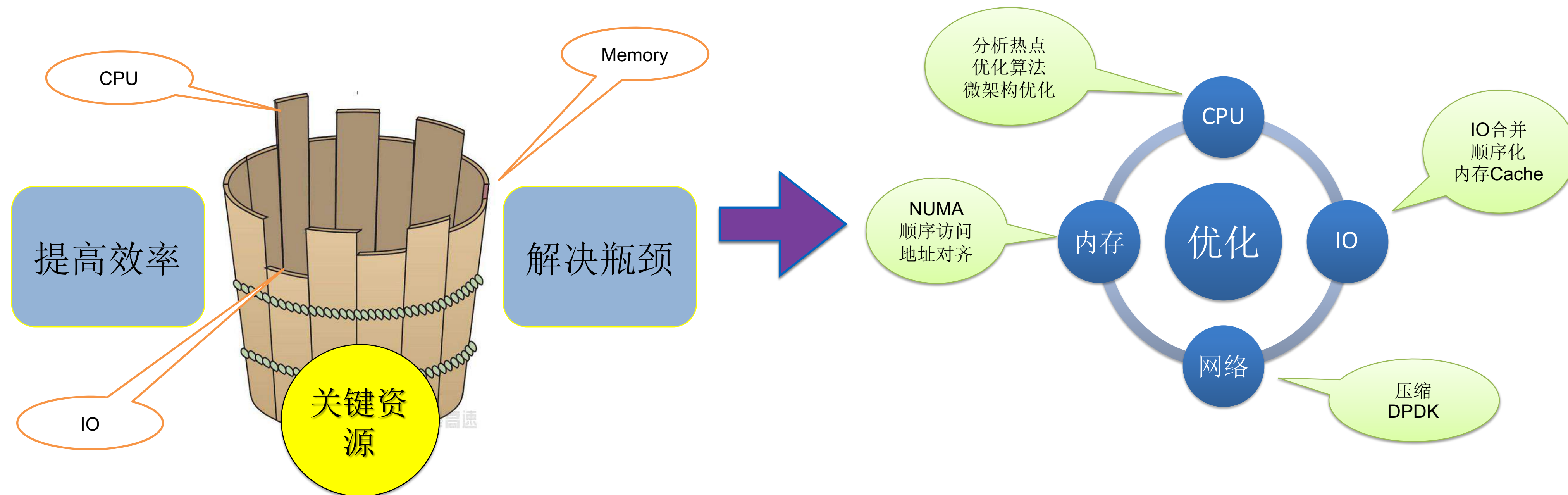
- 核心python package编译成c++ .so lib
- JIT优化
- C++函数指针级别原生调用python
- 轻量级语言安全沙箱



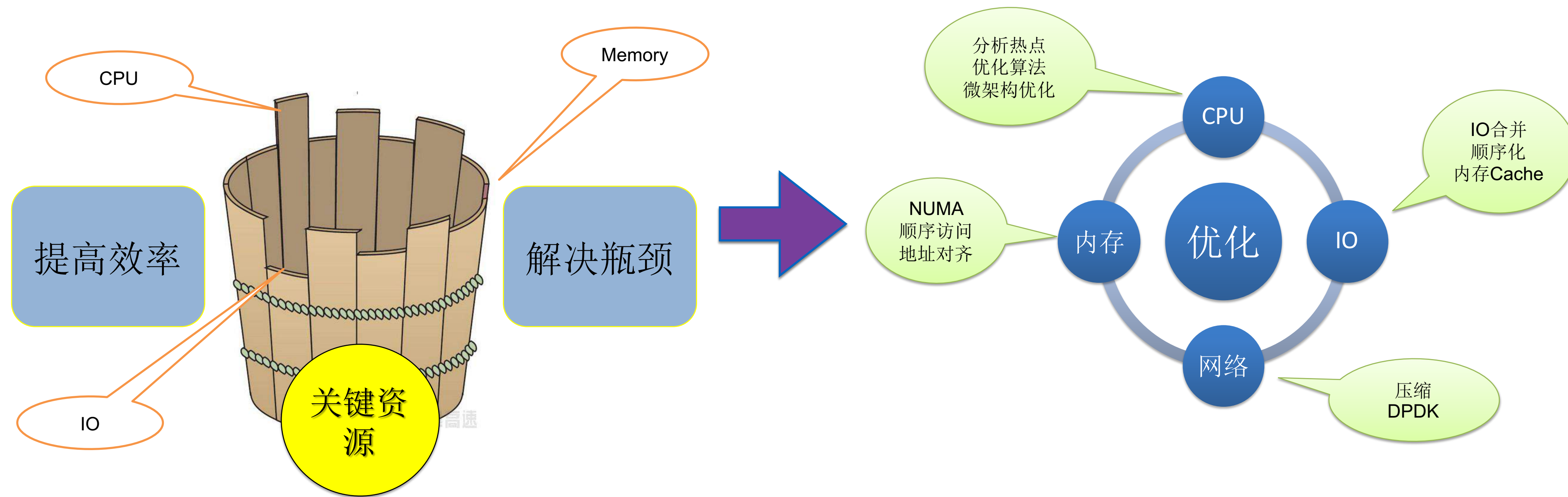
MaxCompute 性能优化——profiling工具 (单机篇)



MaxCompute 性能优化——profiling工具（单机篇）



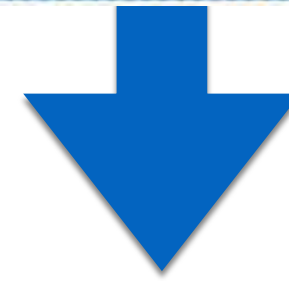
MaxCompute 性能优化——profiling工具（单机篇）



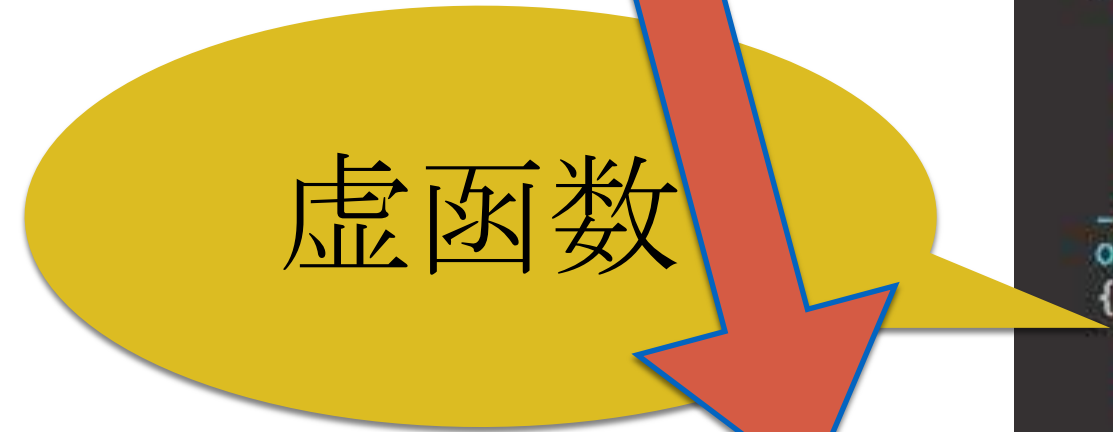
profiling工具：vtune、perf

MaxCompute 性能优化——profiling (案例分析)

apsara::odps::execution_engine::DataRecord::GetSize	856.006ms	1,072,500,000	2.350	1.180
---	-----------	---------------	-------	-------



```
int64_t DataRecord::GetSize()
{
    return BOOL_FLAG(executionengine_EnableDataDump) ? GetFixedSize() + mStringSize : 0;
}
```

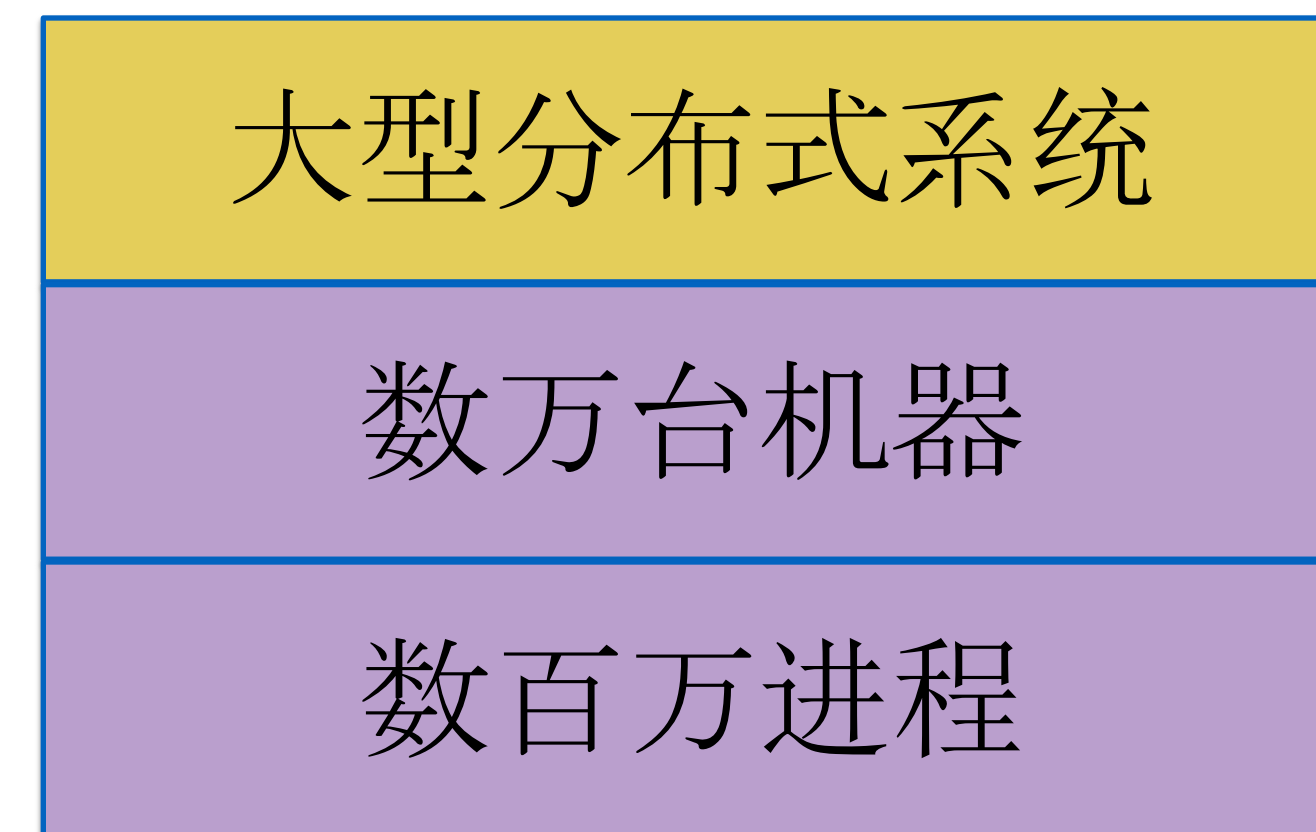


```
0000000000000070 <_ZN6apsara4odps16execution_engine10DataRecord7GetSizeEv>:
DECLARE_FLAG_BOOL(executionengine_EnableDataDump);

int64_t DataRecord::GetSize()
{
    return BOOL_FLAG(executionengine_EnableDataDump) ? GetFixedSize() + mStringSize : 0;
70: 48 8b 05 00 00 00 00    mov     0(%rip),%rax        # 77 <_ZN6apsara4odps16execution_engine10DataRecord7GetSizeEv+0
77: 31 d2                  xor     %edx,%edx
79: 80 38 00              cmpb   $0x0,(%rax)
7c: 74 27                  je     a5 <_ZN6apsara4odps16execution_engine10DataRecord7GetSizeEv+0x35>
_Tp*
operator->() const // never throws
{
    _GLIBCXX_DEBUG_ASSERT(_M_ptr != 0);
    return _M_ptr;
7e: 48 8b 57 10          mov     0x10(%rdi),%rdx
82: 48 8b 42 08          mov     0x8(%rdx),%rax
89: 8b 8a a8 00 00 00    mov     0xab(%rdx),%ecx
8f: 48 c1 f8 03          sar     $0x3,%rax
93: 69 c0 cd cc cc cc    imul   $0xffffffff,%eax,%eax
99: 48 8d 4c 01 28      lea    0x28(%rcx,%rax,1),%rcx
9e: 8b 47 20          mov     0x20(%rdi),%eax
a1: 48 8d 14 01          lea    (%rcx,%rax,1),%rdx
a5: 48 89 d0          mov     %rdx,%rax
a8: c3                retq
a9: 90                nop
aa: 66 0f 1f 44 00 00    nopw  0x0(%rax,%rax,1)
00000000000000b0 <__tcf_14>:
};
```

apsara::odps::execution_engine::DataRecord::GetSize	22.052ms	22,500,000	2.556	1.045
---	----------	------------	-------	-------

MaxCompute 性能优化——profiling工具（集群篇）



传统profiling工具无法进行**job级别**
或者**集群级别**性能分析

MaxCompute 性能优化——profiling工具（集群篇）

扁鹊系统



神医

- 基于云的全站性能分析系统
 - 基于阿里云的多种云服务
 - 全站数据收集、存储、分析、可视化
 - 对目标系统完全无侵入, 无干扰
 - 性能稳定, 开销小
- 与Intel深度合作
 - CPU Profiling、性能优化
- 集群性能分析
 - 热点代码分析
 - 内存使用分析
- 故障诊断
 - Root Cause, 自动、实时诊断