大规模分布式计算

MapReduce 和 Hadoop

徐冬
frankxus@gmail.com

为什么学习大规模分布式计算？

Agenda

- 分布式计算
  - 分布式计算简介
  - MapReduce
  - 应用现状
- MapReduce原理
  - 系统视图
- Hadoop
  - HDFS
  - MapReduce

问题的描述

- 互联网应用的数据量级数
  - Google
    - (2007) 平均每个任务处理180GB
  - Facebook
    - (2010)每日的日志数据超过130TB
  - Taobao
    - (2011) 每日处理超过20TB
  - …

解决思路

- 并行化/分布式
- 通用解决方案

需求

- 大规模数据集
- 成本
- 通用性
- 大规模异构集群的稳定性和容灾能力

MapReduce

- 一种分布式计算的编程模型
- 一套大规模数据处理的通用解决方案
- 一个被互联网公司广泛使用的数据分析基础架构

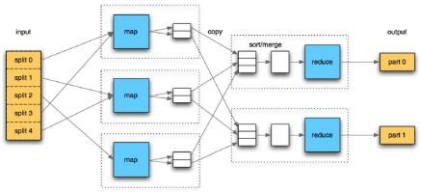
现状

- 应用
  - 大规模日志分析、报表
  - 商业智能、数据挖掘
  - 大规模索引
- 规模

MAPREDUCE

概念

- 一个游戏
- 语义
  - Map：映射，空间变换
  - Reduce：汇总，规约



概念

- 核心理念
  - 任务分片
  - 数据再分布
- <key, value>数据结构
  - map (in_key, in_value) -> list(out_key, intermediate_value)
  - reduce (out_key, list(intermediate_value)) -> list(out_value)
- out_key作为数据再分布的标识

系统视图

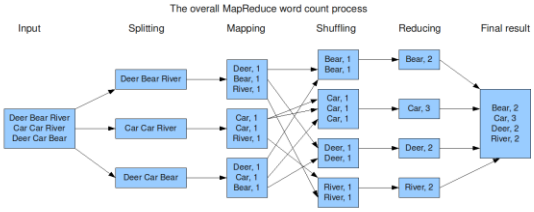
- 容灾能力
  - 单点的Fail-over
- 扩展性
  - Share-nothing
- 并行能力
  - 子任务间并行
  - M/R任务间并行
  - 作业间并行

系统视图

- 任务消耗
  - 网络I/O
  - 磁盘I/O
  - 排序
  - Map/Reduce计算

范例

- 词频统计



范例

- web点击日志分析
  - PV
  - UV/IV…
  - 去重/汇总类需求

范例

- 海量数据排序
  - 外部/归并排序？
  - 分桶
  - 采样

范例

- Join
  - Join key汇总

范例

- 其他范例

应用总结

- MapReduce能？
- MapReduce不能？
  - 在线应用
  - 复杂依赖逻辑（循环、递归？），不可切分的情况？
- MapReduce的数据倾斜问题

HADOOP 介绍

Hadoop简介

- Hadoop
  - 开源的分布式计算系统实现
  - 完整的MapReduce系统栈支持

系统	Hadoop组件	Mimic of?
文件系统（DFS）	HDFS	GFS
MapReduce计算框架	MapReduce	MapReduce
锁服务	ZooKeeper	Chubby
RPC	Avro	ProtocolBuffer*
高级语言/工作流支持	Hive/Pig/Cascading	Sawzaw*
实时（KV）存储	HBase/HyperTable	BigTable

Hadoop简史

- 2004 Google发表有关MapReduce论文
- 2005 Nutch迁移到MapReduce实现
- 2006.1 Doug Cutting加入Yahoo！
- 2006.2 Yahoo从Nutch中剥离出MapReduce并开始使用Hadoop
- 2007.1 Yahoo组建1000+节点的Hadoop集群
- 2008.1 Hadoop成为Apache TLP

设计原则

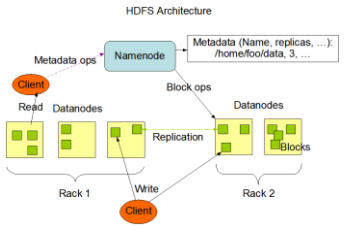
- 硬件错误是常态而不是异常
- 流式数据访问
- 大规模数据集
- 简单的一致性模型
- “移动计算比移动数据更划算”
- 异构软硬件平台间的可移植性

HDFS

- HDFS is
  - 分布式文件系统
  - 支持MapReduce操作（locality支持）
- HDFS is not
  - 不支持文件更新（追加？）
  - 随机读取？

HDFS概念

- 角色
  - NameNode
  - DataNode
  - (DFS)Client
- 数据模型
  - File
  - Block
  - Package
- Rack awareness
- Replication



特性

- 重要特性
  - 大容量/水平扩展能力
  - 容错
  - IO吞吐能力
  - 就近IO策略
- 限制

MapReduce

- 分布式编程框架
- 支持多种编程语言
  - Java
  - C++ (pipes)
  - 其他(streaming)

MapReduce概念

- 角色
  - JobTracker
  - TaskTracker
  - (Job)Client
- 调度模型
  - Queue/Group
  - Job
  - Task
- Scheduler
- (Map side)Locality
- Task failover(attempt)/Speculative execution

编写MapReduce逻辑

- Java API
  - mapper/reducer/combiner
  - partitioner
  - inputformat/outputformat
  - …
- Streaming
- 高级语言实现
  - Hive
  - Pig