Apache CarbonData成长故事

China Open Source Conference 2017

演讲者： 陈亮

目录

CONTENTS

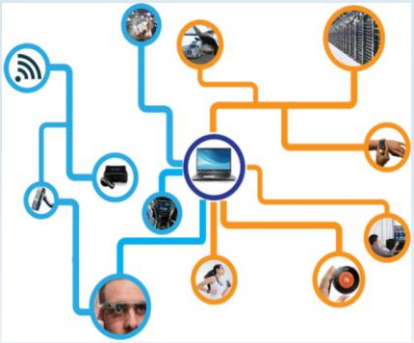
项目背景

华为开启CarbonData项目的背景是什么？
这个项目的价值是什么？

Big Data Challenges in Huawei(1)

Network

- 54B records per day
- 750TB per month
- Complex correlated data



Consumer

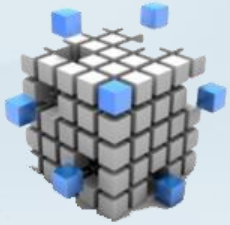
- 100 thousands of sensors
- >2 million events per second
- Time series, geospatial data

Enterprise

- 100 GB to TB per day
- Data across different domains

Big Data Challenges in Huawei(2)

Report & Dashboard



OLAP & Ad-hoc

Batch processing

Machine learning

Realtime Decision

data

Challenge Summary - Data

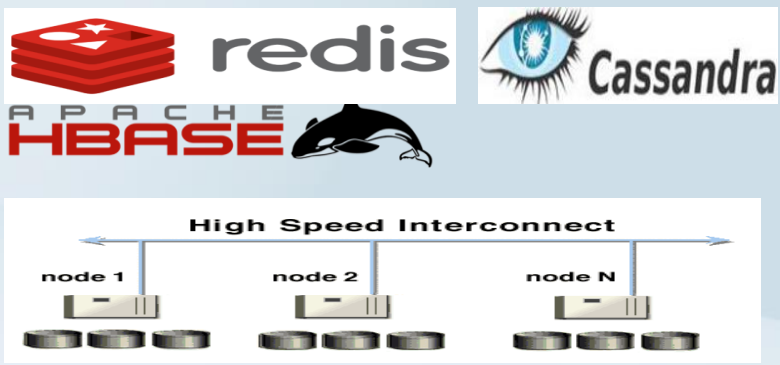
- Data Size 数据规模
  - Single Table >10 B　单表大于100亿行
  - Fast growing　　快速增长
  - Nested data structure for complex object　数据结构复杂

- Multi-dimensional 数据维度多
  - Every record > 100 dimensions　　分析的维度超过100
  - Add new dimension occasionally　维度不断增长
  - Billion level high cardinality　　不同值范围在亿级别

当前大数据开源技术：无法满足一份数据同时支撑多种大数据场景的需求

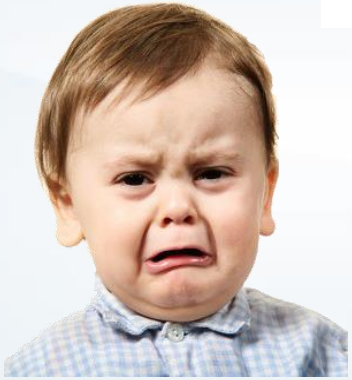
1. NoSQL Database
   - 只支持单列key value查询 <5ms
   - 不支持标准SQL
2. MPP relational Database
   - Shared-nothing架构
   - 不支持大集群 <100节点，没有容错
3. Cube Data
   - 预聚合，查询快
   - 但数据膨胀大，支持维度少，不支持查明细数据
4. Search Engine
   - 通过索引快速找到数据
   - 数据膨胀大2-4倍，不支持SQL
5. SQL on Hadoop
   - 聚焦计算引擎的分布式扫描
   - 存储效率不高

redis Cassandra HBASE



Lucene
Apache Solr
Elasticsearch

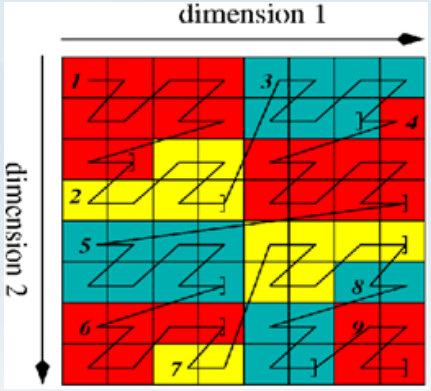
APACHE DRILL HAWQ presto HIVE



互联网场景 ——>针对某些场景的大数据方案

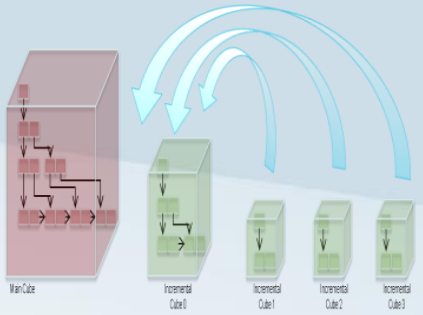
如何通过一份数据高效支持多种业务场景诉求，减少数据孤岛和冗余？

快：
构建多维度索引



更快：
基于快速智能扫描

高效数据压缩：
字典编码



并发数据导入：
Spark并行任务

- CarbonData名字由来。
- 经过几年的技术开发，CarbonData性能提升X-XX倍。

开源准备

公司流程+开源准备工作

开源准备checklist：

- 法务品牌审视
- 开源基金会分析
- 公司汇报
- 组建开源发展团队(OSDT)

选基金会：Apache基金会和Linux基金会的对比

基金会运作	税务结构	资助形式	雇员	参与模式	运作范围
Apache	501(c)(3)	捐赠	很少	个人	项目
Linux	501(c)(6)	会员	较多	企业	项目+会议+培训+咨询

项目运作	项目范围	项目个数	允许相似项目	社区导向	公司话语权	项目决定权
Apache	大数据、数据库、Web、云服务、AI	300+	是	纯技术，真正的开源社区->项目多，生态好	零	1.Board季度review项目运作情况，为项目提供INFRA、法务等保障 2.项目自主管理diversity(项目方向+核心人员)
Linux	云计算(INFRA)、OS、网络	60+	否	技术+商业->组织化运作	代表公司利益	董事会+项目

CarbonData为什么要开源？

- 将CarbonData打造为标准和通用数据格式/存储
- 数据无Lock-in安全问题，客户用得放心
- 构建CarbonData生态
- 提升开源影响力
- 提升开发效率

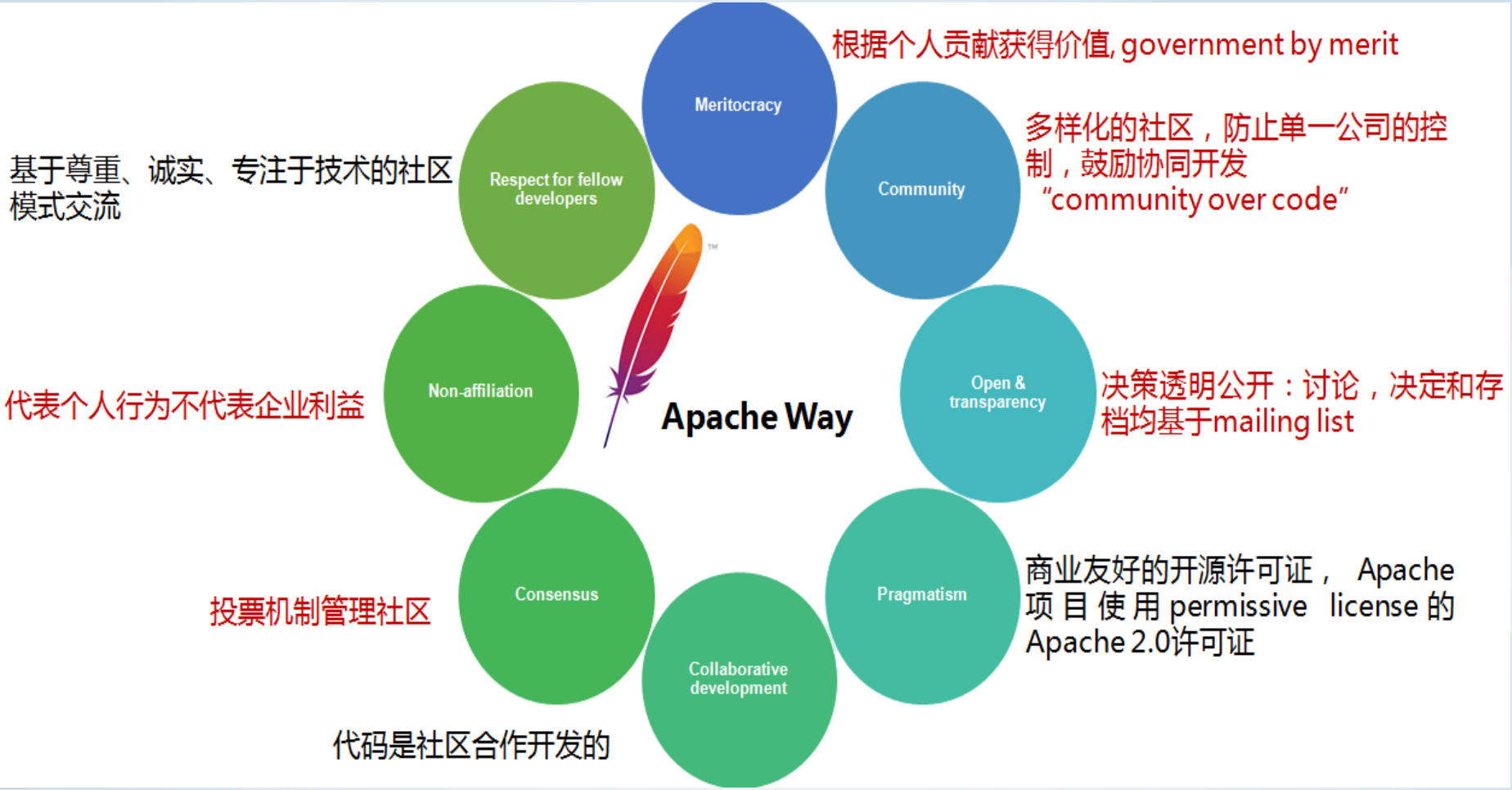
…

做大蛋糕，比做大份额更重要！

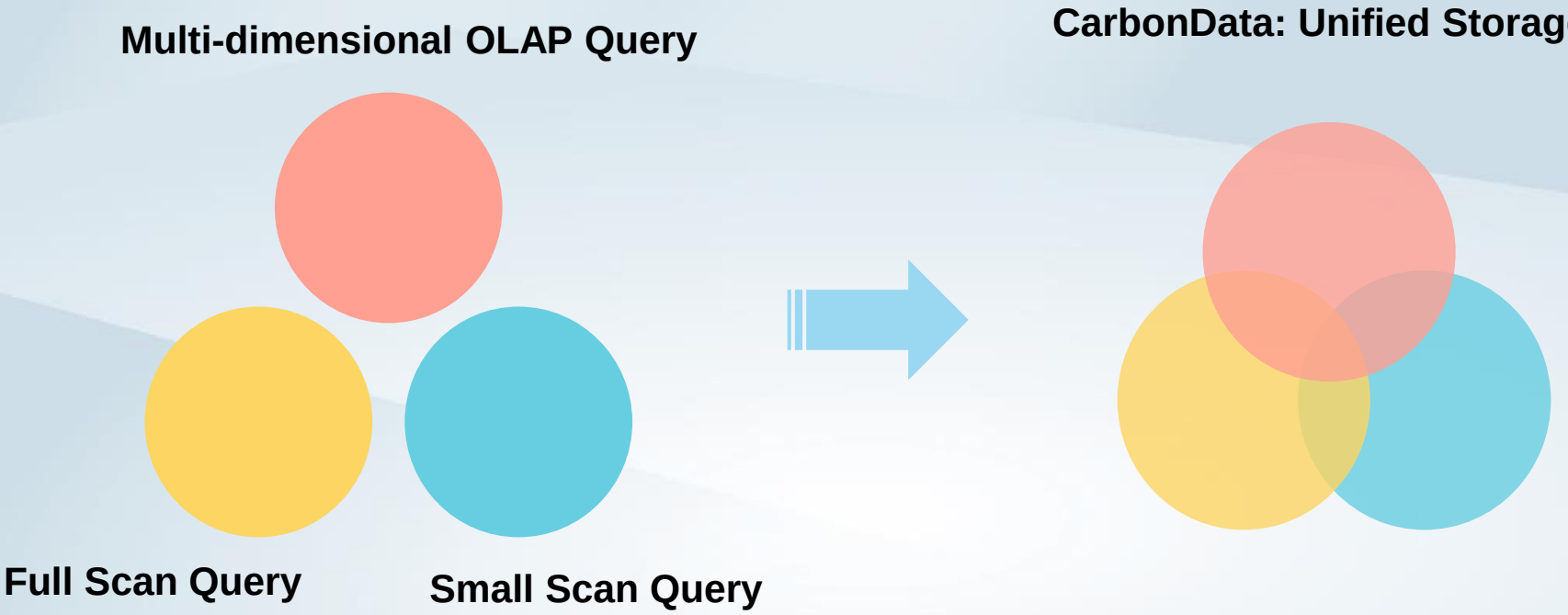
从Apache孵化到毕业的历程

Apache项目生态构建关键路径分析

“Apache Way”是ASF的精髓



CarbonData：实现一份数据同时满足多种业务需求，与大数据生态无缝集成



CarbonData官网：carbondata.apache.org
Github : https://github.com/apache/carbondata

感谢您的聆听

演讲者：陈亮
微　信：chenliang2007
Email： chenliang613@apache.org