Apache CarbonData

陈亮 Apache PMC,Committer
(chenliang613@apache.org)

oschina号：HW陈亮

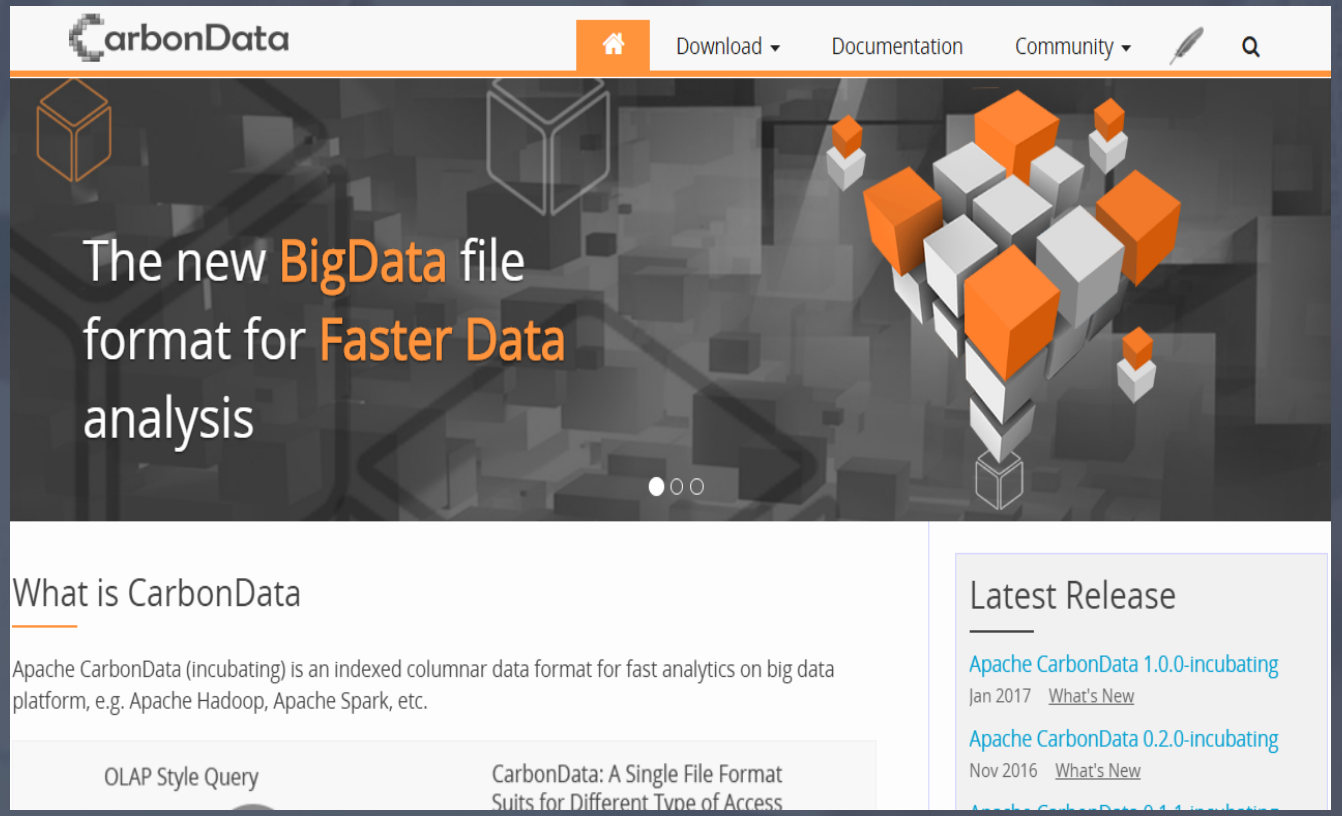
擅长的领域： 大数据

公司的行业背景：华为

码云地址：
https://git.oschina.net/CarbonData/ApacheCarbonData

项目简介

Apache CarbonData是一种新的高性能数据存储格式，针对当前大数据领域分析场景需求各异而导致的存储冗余问题，CarbonData提供了一种新的融合数据存储方案，以一份数据同时支持“任意维度组合的过滤查询、快速扫描、详单查询等”多种应用场景，并通过多级索引、字典编码、列存等特性提升了IO扫描和计算性能，实现百亿数据级秒级响应。



(http://carbondata.apache.org)

当前大数据的各种挑战

- 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

当前各种大数据方案分析

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

APACHE HBASE

Lucene

Apache Solr

Elasticsearch

APACHE DRILL

HAWQ

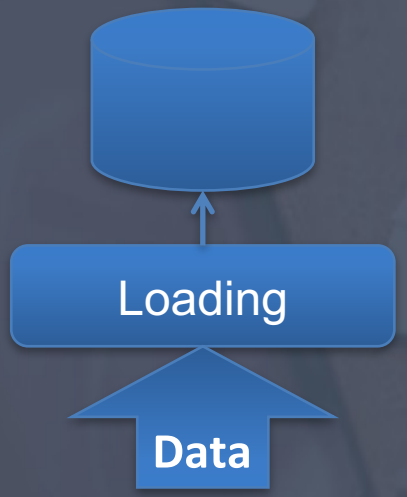
presto

HIVE

架构师的苦恼：不同应用不同数据存储，如果

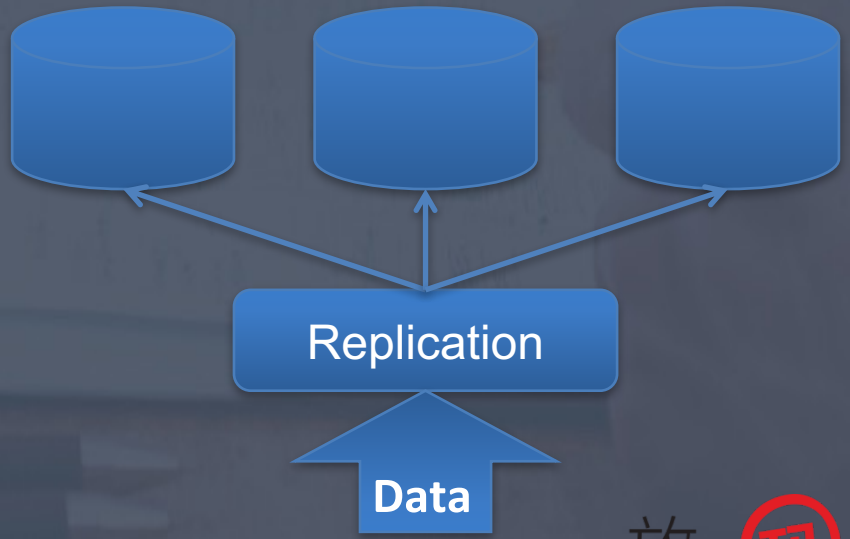
Choice 1: Compromising

App1 App2 App3

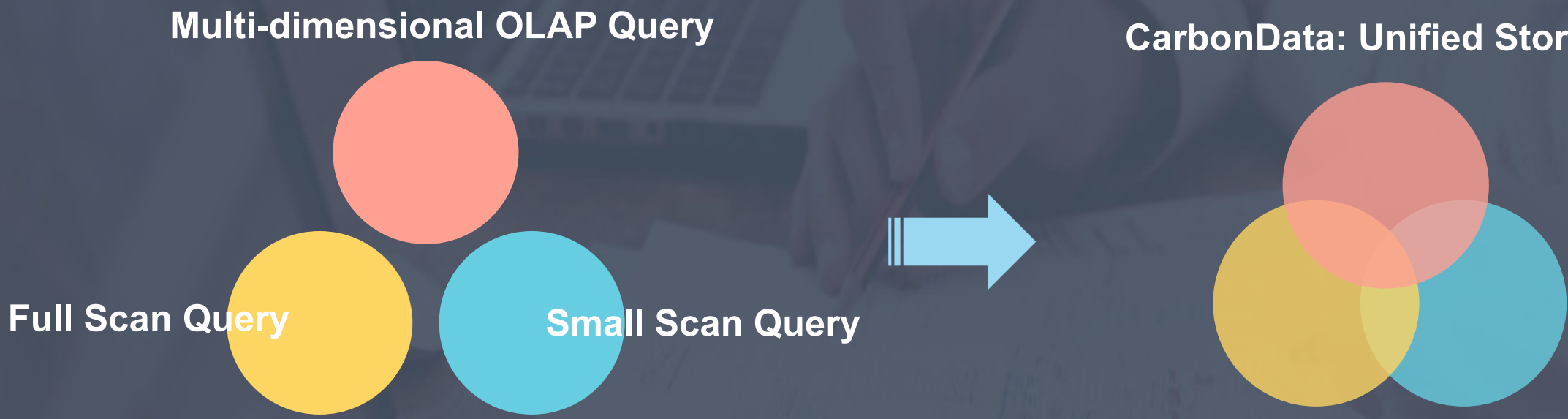


Choice 2: Replicating of data

App1 App2 App3

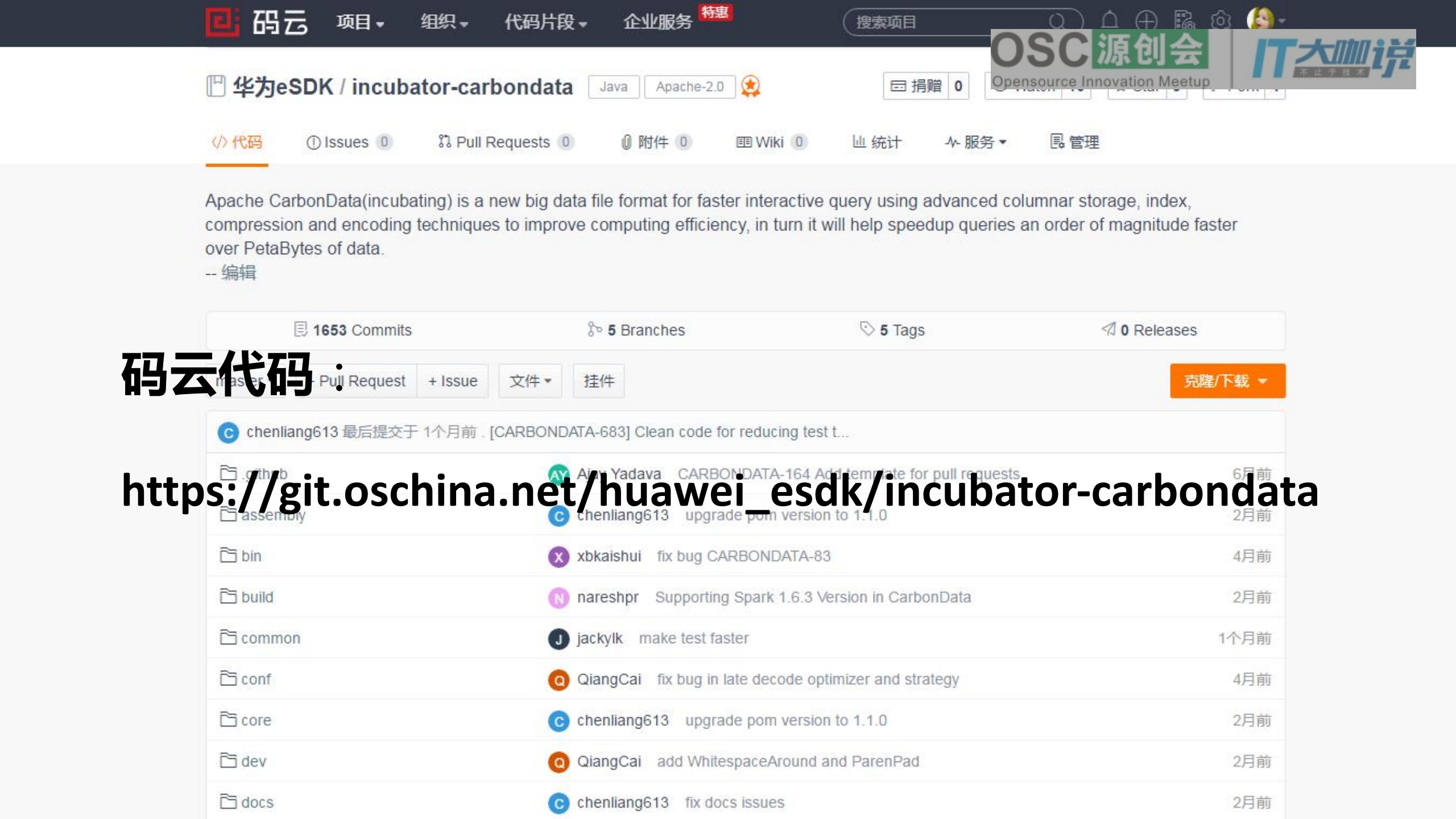


CarbonData：实现一份数据同时满足多种业务场景，大数据生态无缝集成



独特的特性：

- 各种索引(多维索引，倒排索引，MinMax索引)
- 字典编码
- 数据更新、删除
- 大数据生态无缝集成(Hadoop,Spark)



码云代码：

https://git.oschina.net/huawei_esdk/incubator-carbondata

Apache CarbonData社区

- CarbonData 2016年6月全票通过正式进入Apache基金会.
- 已发布了4个Apache稳定版本
- Welcome contribution:
  - **码云代码：https://git.oschina.net/huawei_esdk/incubator-carbondata**
  - Apache github代码： https://github.com/apache/incubator-carbondata
  - website: http://carbondata.apache.org
  - JIRA: https://issues.apache.org/jira/browse/CARBONDATA
  - Mail list: dev@carbondata.incubator.apache.org
- 用户：Inmobi, 美团, 蚂蚁金服, Hulu, 滴滴，交行, 建行, 金陵科技, 上汽等

OSC 源创会 Opensource Innovation Meetup | IT大咖说 不止于技术

THANK YOU!

如果大家感兴趣Apache CarbonData项目，可以添加我的微信：chenliang2007，我邀请各位加入Apache CarbonData User Group微信群(因超过100人，只能邀请)

http://carbondata.apache.org

放码过来