

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

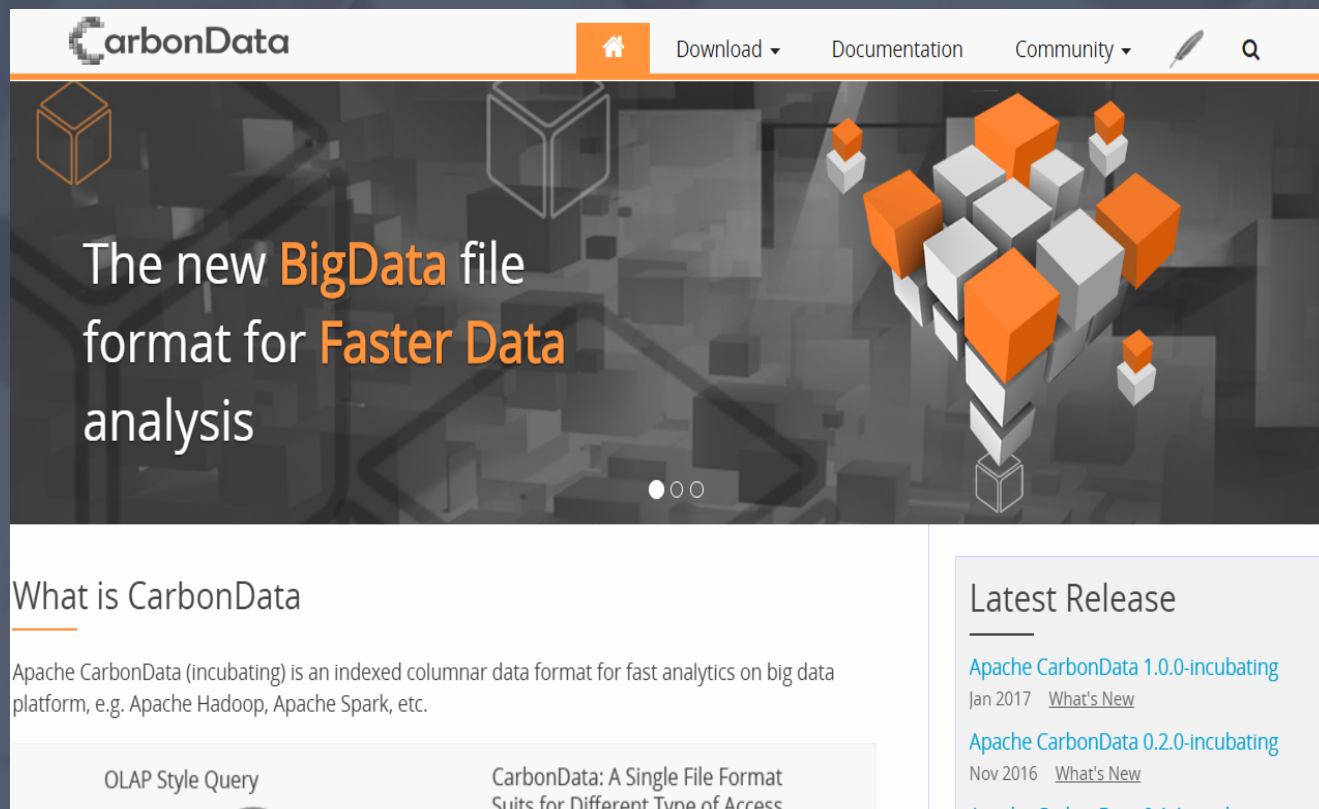
码云地址：

<https://git.oschina.net/CarbonData/ApacheCarbonData>

放  过 来

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://carbonda.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

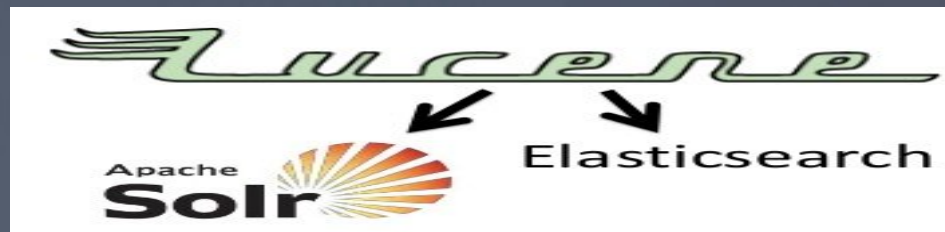
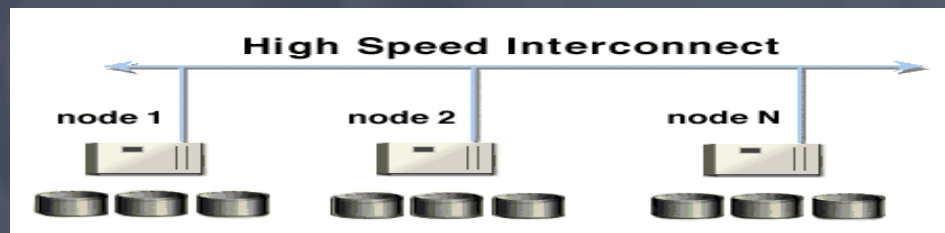
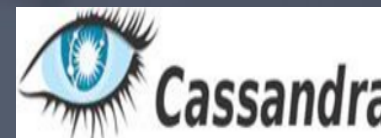
- 预聚合，查询快
- 但数据膨胀大，支持维度少，不支持查明细数据

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

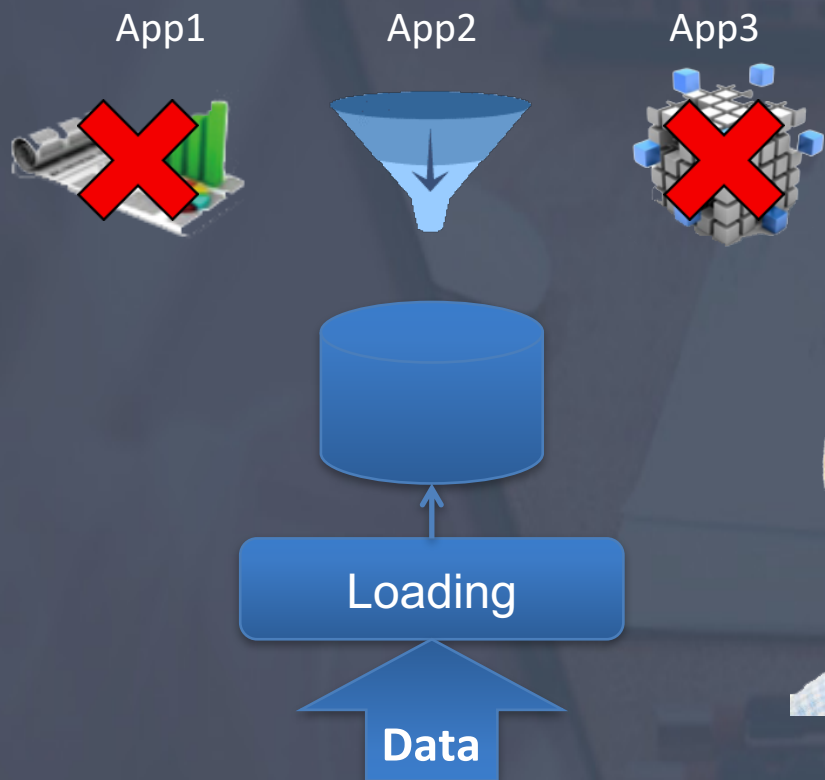
5. SQL on Hadoop

- 聚焦计算引擎的分布式扫描
- 存储效率不高

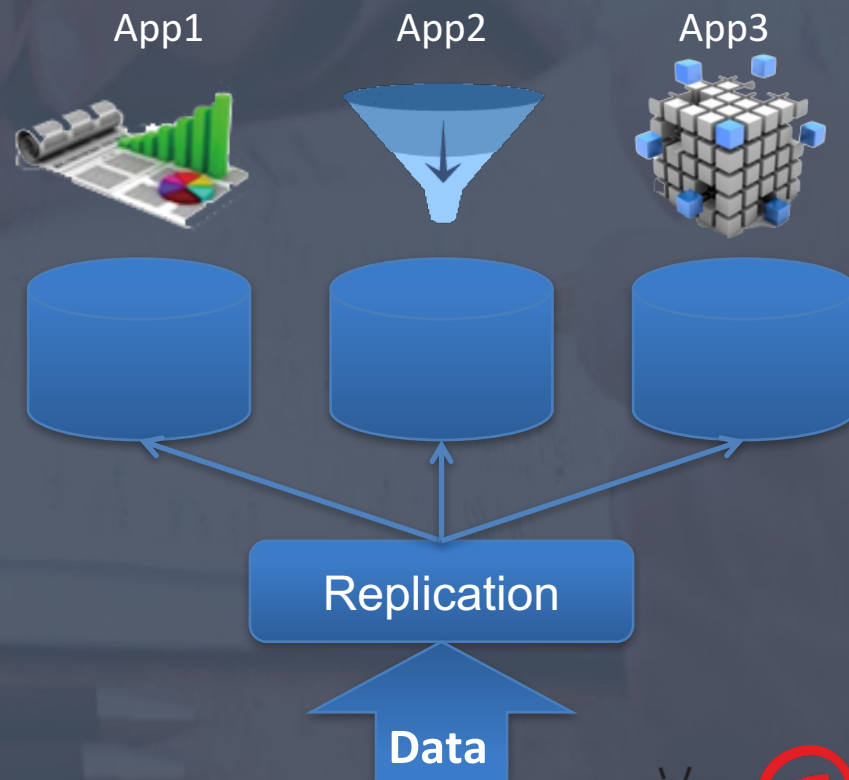


架构师的苦恼：不同应用不同数据存储，如果

Choice 1: Compromising



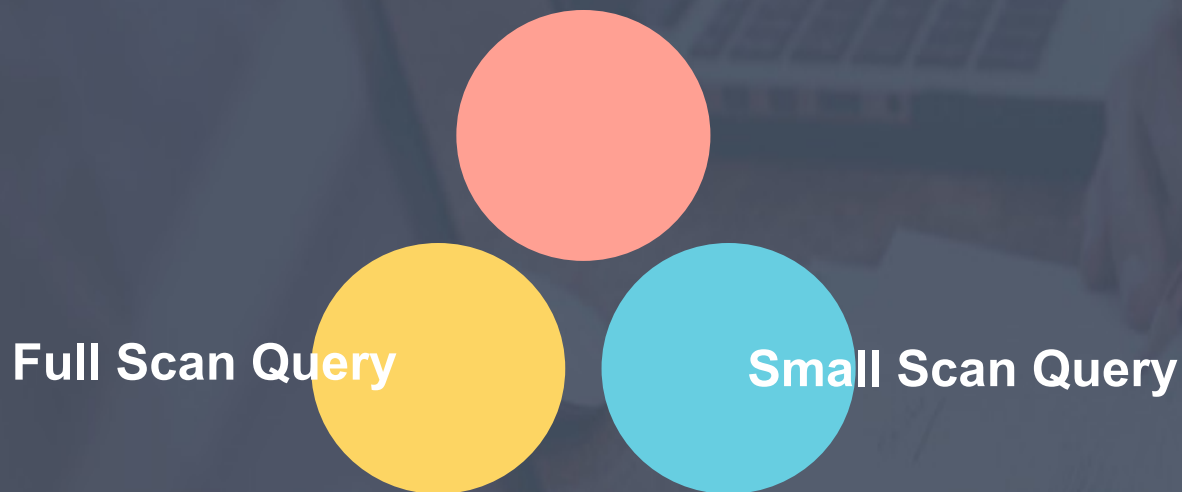
Choice 2: Replicating of data



放  过来

CarbonData : 实现一份数据同时满足多种业态无缝集成

Multi-dimensional OLAP Query



CarbonData: Unified Storage



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)

放  过 来

华为eSDK / incubator-carbondata Java Apache-2.0

捐赠 0

代码 Issues 0 Pull Requests 0 附件 0 Wiki 0 统计 服务 管理

Apache CarbonData(incubating) is a new big data file format for faster interactive query using advanced columnar storage, index, compression and encoding techniques to improve computing efficiency, in turn it will help speedup queries an order of magnitude faster over PetaBytes of data.

-- 编辑

1653 Commits 5 Branches 5 Tags 0 Releases

码云代码：

Master Pull Request + Issue 文件 挂件 克隆/下载

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

chenliang613 最后提交于 1个月前 . [CARBONDATA-683] Clean code for reducing test t...		
.github	Aj Y Yadava CARBONDATA-164 Add template for pull requests	6月前
assembly	chenliang613 upgrade pom version to 1.1.0	2月前
bin	xbkaishui fix bug CARBONDATA-83	4月前
build	nareshpr Supporting Spark 1.6.3 Version in CarbonData	2月前
common	jackylk make test faster	1个月前
conf	QiangCai fix bug in late decode optimizer and strategy	4月前
core	chenliang613 upgrade pom version to 1.1.0	2月前
dev	QiangCai add WhitespaceAround and ParenPad	2月前
docs	chenliang613 fix docs issues	2月前

Apache CarbonData社

OSC 源创会
OpenSource Innovation Meetup

IT大咖说
不止于技术

- 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

放  过 来

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  过 来