Spinach: 基于Spark SQL在生产环境实现即席查询

王道远(Intel)，李元健 (百度)

Notice and Disclaimers:

- Intel, the Intel logo are trademarks of Intel Corporation in the U.S. and/or other countries. *Other names and brands may be claimed as the property of others.
  See Trademarks on intel.com for full list of Intel trademarks.
- Optimization Notice:
  Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel.
  Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.
- Intel technologies may require enabled hardware, specific software, or services activation. Check with your system manufacturer or retailer.
- No computer system can be absolutely secure. Intel does not assume any liability for lost or stolen data or systems or any damages resulting from such losses.
- You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.
- No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.
- The products described may contain design defects or errors known as errata which may cause the product to deviate from publish.

自我介绍

- 英特尔亚太研发有限公司资深软件工程师
- 我们团队专注于Apache Spark的优化
- 自2014年起就是Spark的活跃贡献者之一，主要集中在Spark SQL模块中
- 译有《Spark快速大数据分析》一书

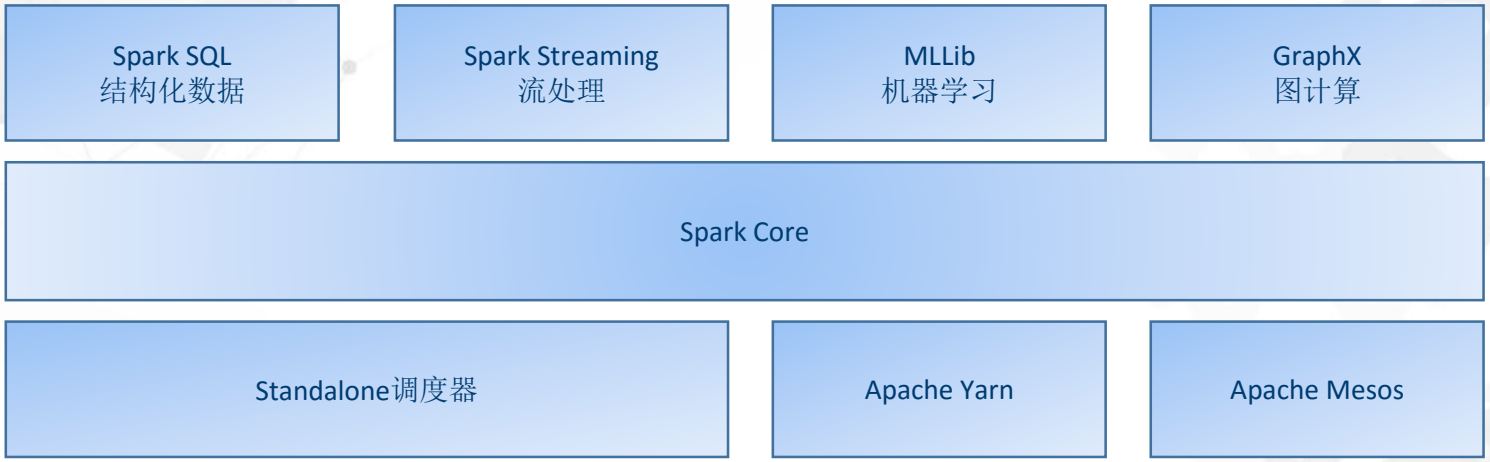
Agenda

- 背景介绍
- Spinach介绍
- Spark和Spinach在百度
  - Spark在百度
  - 百度BigSQL
  - BigSQL中的优化
- 未来计划

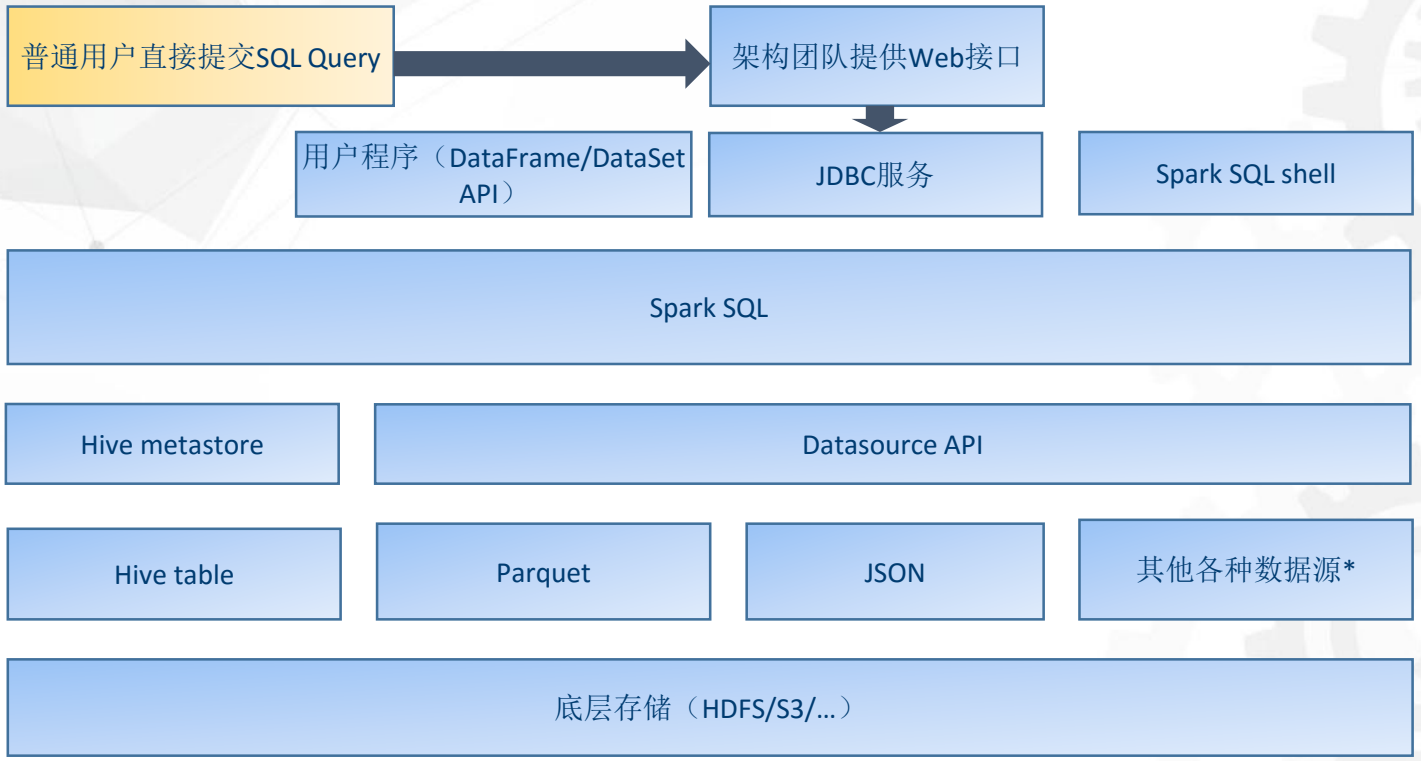
Agenda

- 背景介绍
- Spinach介绍
- Spark和Spinach在百度
  - Spark在百度
  - 百度BigSQL
  - BigSQL中的优化
- 未来计划

关于Apache Spark



使用Spark SQL


普通用户直接提交SQL Query
架构团队提供Web接口
用户程序（DataFrame/DataSet API）
JDBC服务
Spark SQL shell
Spark SQL
Hive metastore
Datasource API
Hive table
Parquet
JSON
其他各种数据源*
底层存储（HDFS/S3/...）

如何优化这种场景？

- 大规模数据集上的数据查询还不够快
- 定时任务式的作业更新结果不够及时

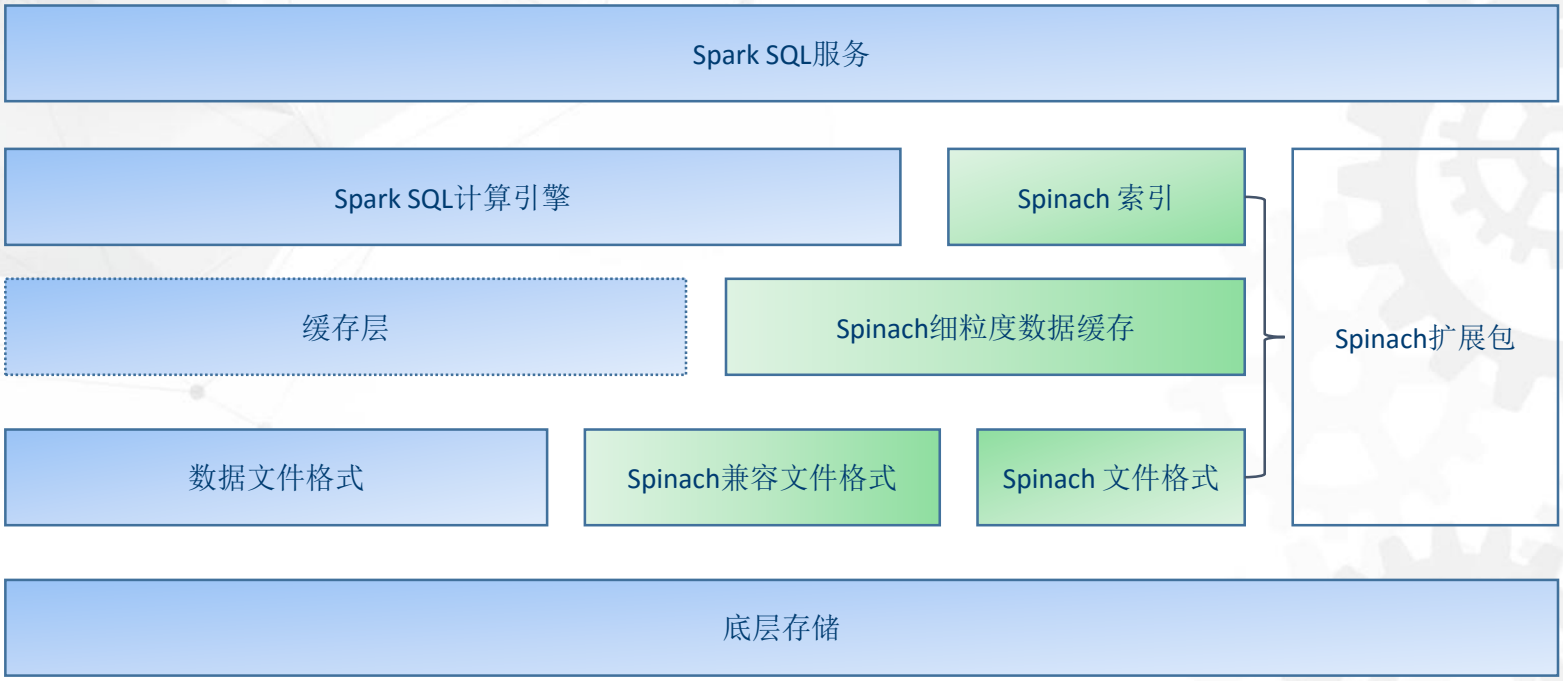
Spark是为通用计算而设计的分布式计算引擎，针对即席查询，还可以进一步优化

如何让Spark SQL做到更快？

Agenda

- 背景介绍
- Spinach介绍
- Spark和Spinach在百度
  - Spark在百度
  - 百度BigSQL
  - BigSQL中的优化
- 未来计划

开源解决方案：Spinach



一个简单的例子

1. 引入Spinach

```
$SPARK_HOME/sbin/start-thriftserver --package spinach.jar
```

2. 创建一张spinach格式的表

```
spark-sql> CREATE TABLE src(a: Int, b: String) USING spn;
```

3. 创建一个单列的B+树索引

```
spark-sql> CREATE SINDEX idx_1 ON src (a) USING BTREE;
```

4. 和寻常一样插入数据

```
spark-sql> INSERT INTO TABLE src SELECT key, value FROM xxx;
```

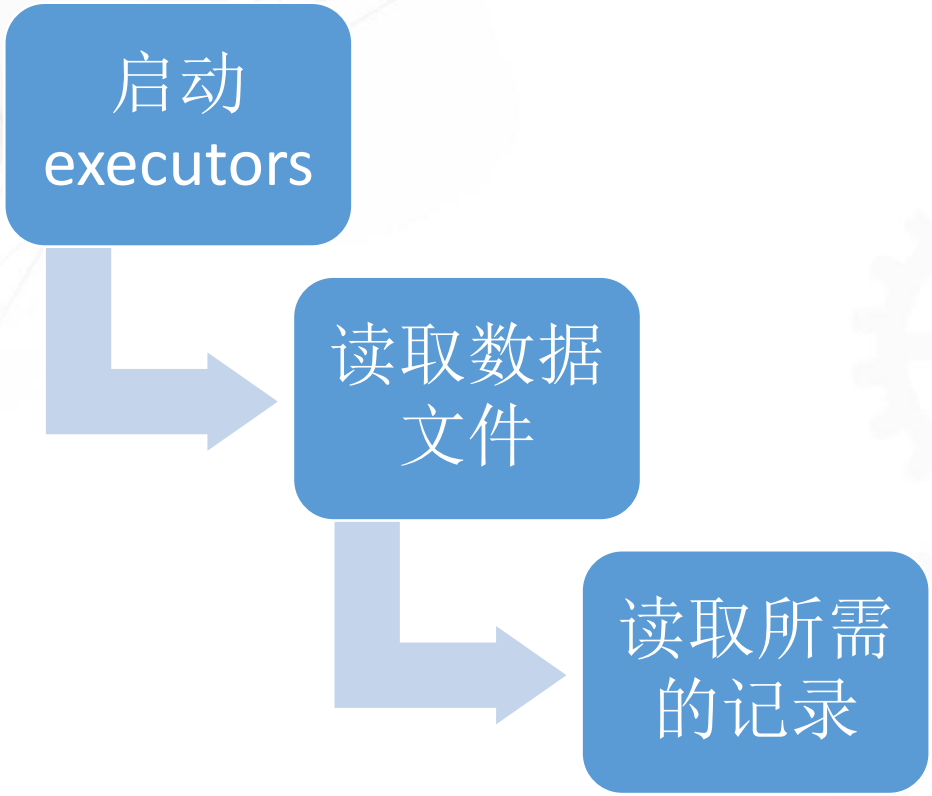
5. 刷新索引

```
spark-sql> REFRESH SINDEX on src;
```

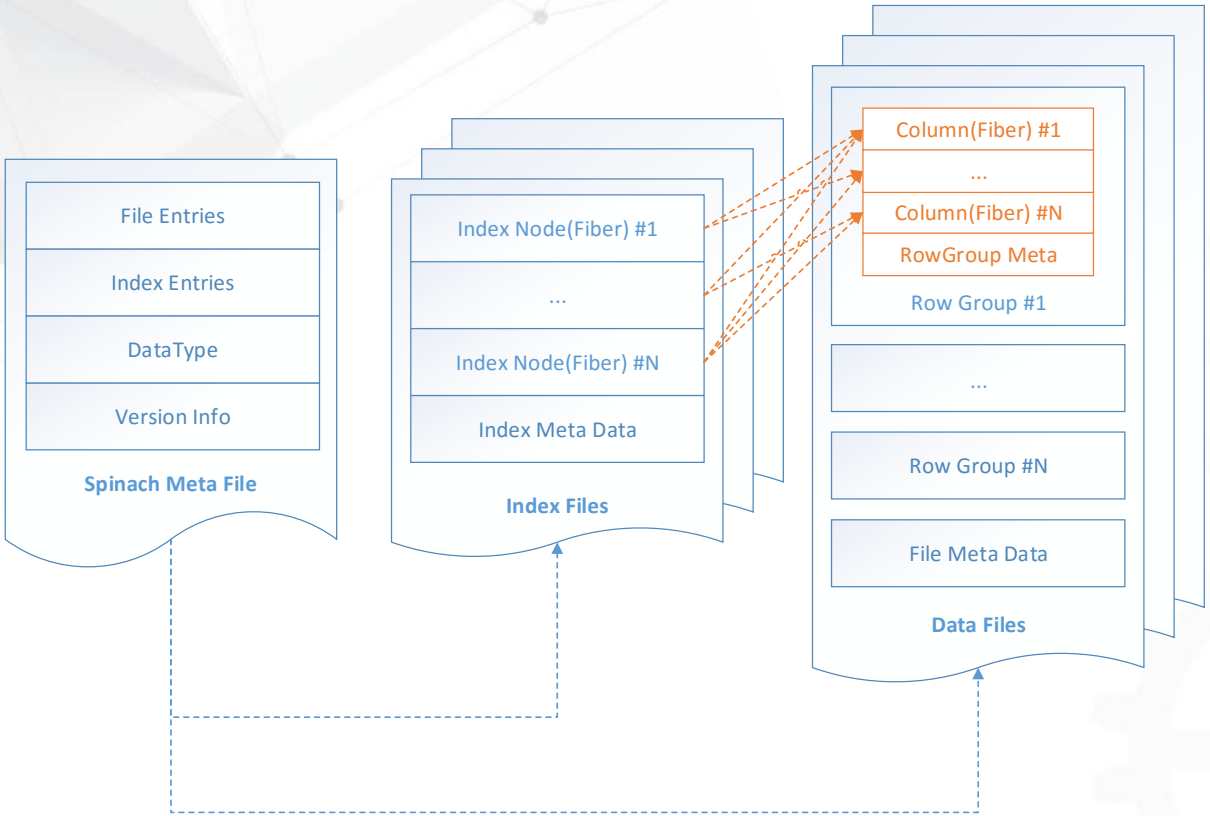
6. 执行查询，查询会被索引自动优化

```
spark-sql> SELECT MAX(value), MIN(value) FROM src WHERE a > 100 and a < 1000;
```

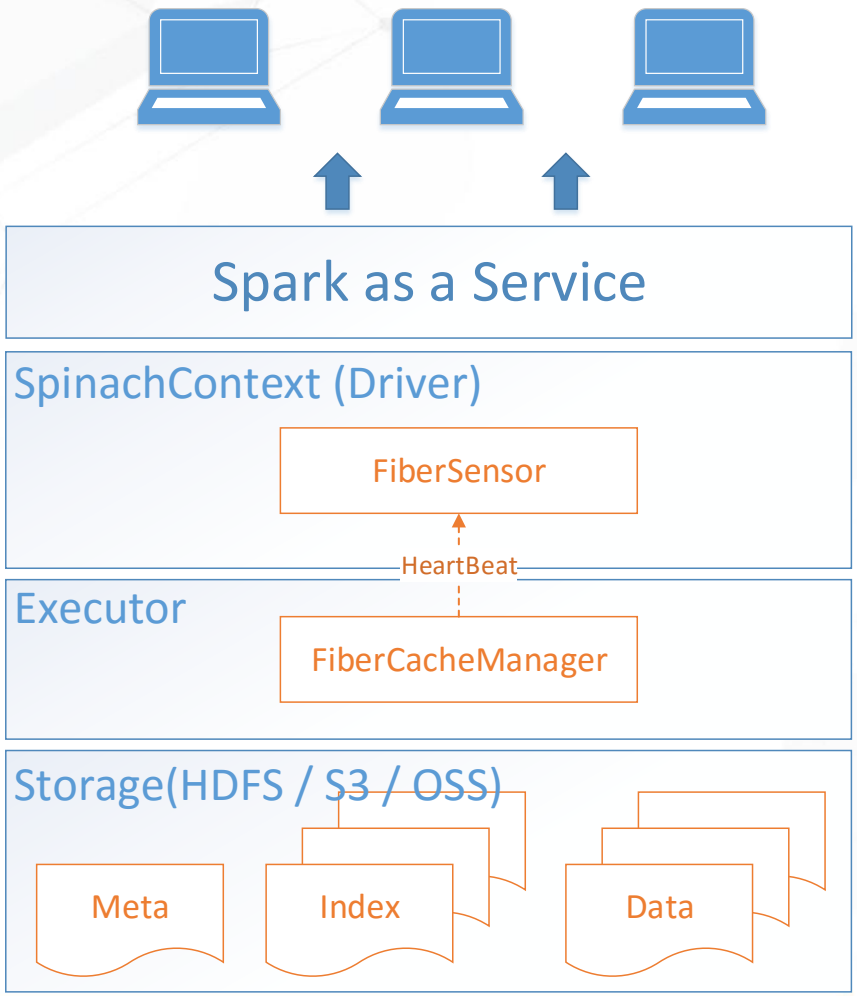
在Spark SQL上运行即席查询



Spinach文件格式结构



Spinach运行时架构



Spinach索引

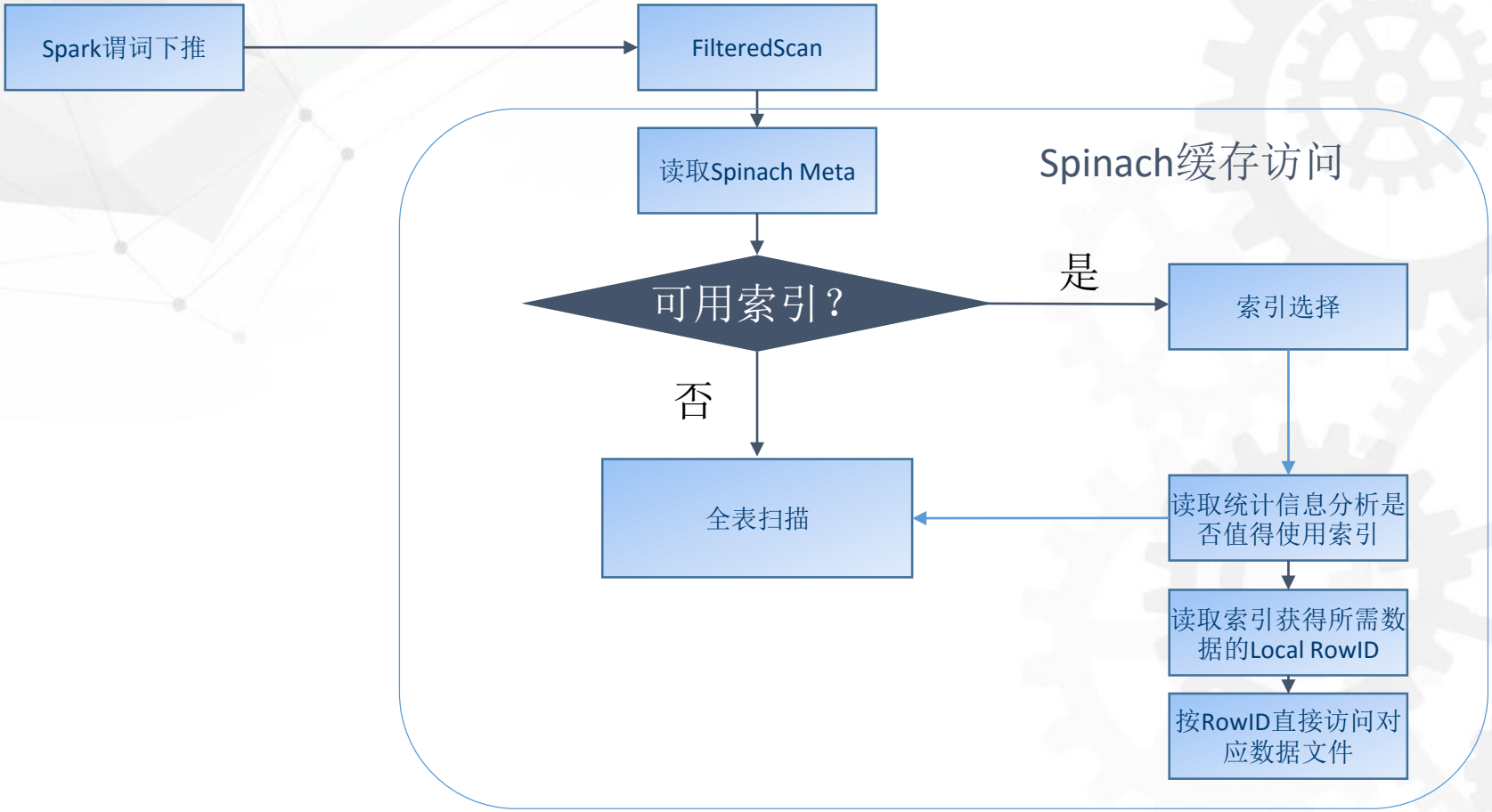
B+树索引

- 支持等值查找与范围查找
- 适用于取值范围广、数据随机的列

Bitmap索引

- 支持等值查找
- 适用于值域不大的列

Spinach加速原理



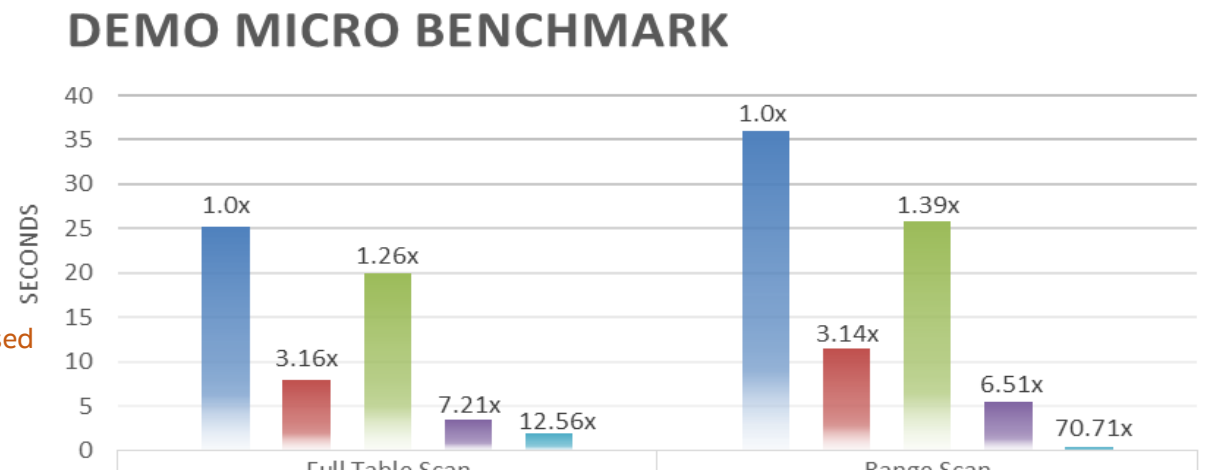
Spinach的优势

成本低	效果好	简单易用
• 充分利用现有硬件环境 • 开源软件	• 类似传统数据库的索引 • 实测5倍性能提升	• 部署简单 • 维护方便 • 符合用户使用习惯

Simple Cache + Index Benchmark

Hardware
- 1 Master + 3 Slaves
- Xeon E5-2699 v3
- 256GB / node
- 8x1TB SATA / node
- 10Gb network

Data
- ~100GB (Uncompressed Parquet*)
- Record Count: 1.3B



	Full Table Scan	Range Scan
Parquet* (Compressed Cold Data)	25.25	36.06
Parquet* (Compressed)(OS Cached Data)	8	11.5
Parquet* (Non-Compressed) (OS Cached)	20	25.9
Spark* In Memory Cached Table	3.5	5.54
Adhoc Query Engine Table	2.01	0.51

- 全表扫描:
  - df.selectExpr("count(str1)", "count(int1)", "count(str2)").show
- 键值范围查询:
  - df.filter("str2 >= 'China-6234567' and str2 <= 'China-6234596'").selectExpr("count(str1)", "sum(int1)").show
- 测试结果可能因数据集特征、操作系统、硬件、软件配置不同而有所不同*

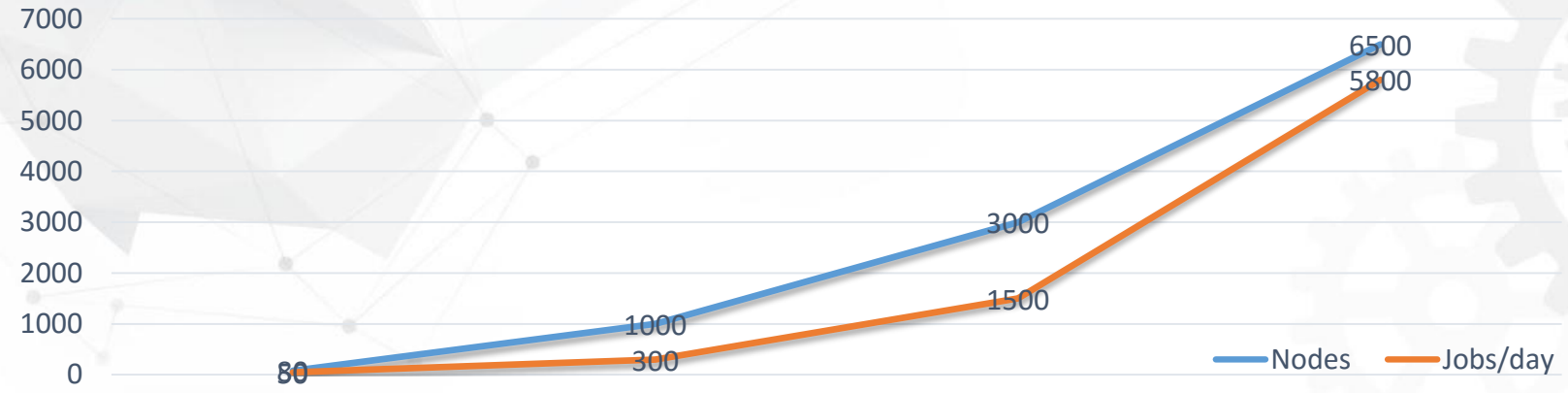
Agenda

- 背景介绍
- Spinach介绍
- Spark和Spinach在百度
  - Spark在百度
  - 百度BigSQL
  - BigSQL中的优化
- 未来计划

自我介绍

- 百度基础架构部 分布式计算方向
- Apache Spark开源贡献者
- Baidu Spark 团队负责人

百度与Spark



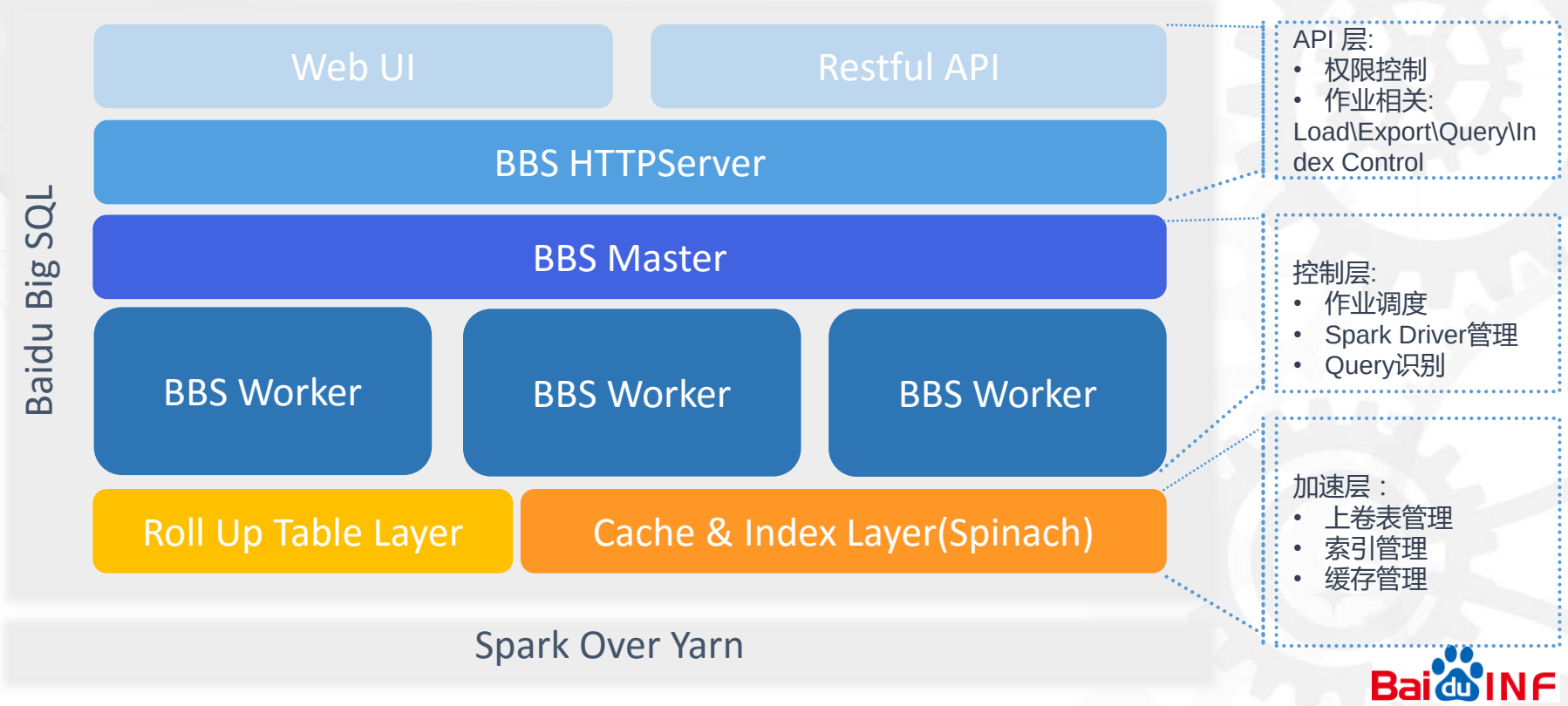
- 引入开源版本
- Version: 0.8

- 构建专用standalone集群
- 完成与公司自研存储、实时、数仓系统对接
- Version: 1.4

- 与公司自研调度系统打通，构建大规模集群
- Version: 1.6

- 构建基于Spark的独立服务
- Spinach
- Version: 2.1

百度Big SQL


Baidu Big SQL
Web UI
Restful API
BBS HTTPServer
BBS Master
BBS Worker
BBS Worker
BBS Worker
Roll Up Table Layer
Cache & Index Layer(Spinach)
Spark Over Yarn
API 层:
• 权限控制
• 作业相关:
Load\Export\Query\Index Control
控制层:
• 作业调度
• Spark Driver管理
• Query识别
加速层：
• 上卷表管理
• 索引管理
• 缓存管理

实际应用


Baidu百度
鲜花
百度一下
百度首页 消息 设置 xuanyuanking
网页 新闻 贴吧 知道 音乐 图片 视频 地图 文库 更多»
百度为您找到相关结果约100,000,000个
搜索工具
鲜花 3小时鲜花 优选中国鲜花网
鲜花,中国鲜花网-国内优秀鲜花服务商,中国花卉协会单位,24小时鲜花,1-3小时送达全国600多城市.国内鲜花优秀品牌!诚信经营,订..
价格区间: 1-100元 | 100-200 热门品牌: 红玫瑰 | 郁金香
www.xianhua.com.cn 2017-05 - 评价 - 广告
鲜花 花礼网-中国鲜花礼品网 1-3小时送花服务
鲜花--花礼网,销量连续5年全国领先,1-3小时送达全国1000多城市.鲜花认证行业龙头企业;诚信经营,用心服务,打造品牌百年老店!
www.hua.com 2017-05 - 1014条评价 - 广告
鲜花roseonly一生只爱一人 爱只送roseonly
roseonly矜贵玫瑰,新娘婚纱花束,高贵鲜花礼盒 只为将"一生只爱一人"完美传达.对至爱的深情告白,只选roseonly
www.roseonly.com.cn 2017-05 - 评价 - 广告
鲜花-野兽派订鲜花 勇敢爱
鲜花-野兽派订鲜花免运费,送货时间可定制,让你的爱不延迟,高端定制俘获她的心!更为您准备了香氛,美妆,家纺等多种精心礼物,替您表达爱!
www.thebeastshop.com 2017-05 - 评价 - 广告
观花植物的品种大全 (432个品种)
花期: 春季 夏季 秋季 冬季 全年
颜色: 蓝紫色 白色 黄色 红色 粉红色 紫红色 橙色 绿色
花礼网，送花就上hua.com!
HUA
5.14 母亲节
鲜花早订更优惠
5.11-5.15
花礼网成立于2005年，11年鲜花品牌服务商。鲜花订单送前实拍保证效果，1-3小时送达鲜花!
【鲜花】24小时订花，配送全国1000城市
【优惠】5.14母亲节，更多节日折扣优惠
http://www.hua.com/ - 品牌广告
登录百度账户 交易更有保障
相关植物
展开
四大名花
牡丹菊花山
杰拉尔顿腊花
红运当头
冬春季室内
蓝色妖姬
寓意清纯敦

实际应用

找到为搜索词“鲜花”付费最多的前十名用户

数据分析 | 主题分析 HOT | 例行化 | 系统升级

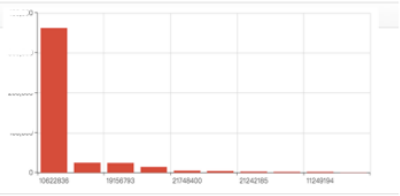
```
1  --- 鼠标移出输入框后，将自动检测可查询
2  select userid, sum(charge) as charge
3  from [illegible]
4  where event_day=20170104
5  and query = '鲜花'
6  group by userid
7  order by charge desc
8  limit 10
```

任务名称，默认result_时间 | 执行 | 清空 | 上传词表 | 上传UDF

任务	SQL语句	任务状态
任务编号：201405 JobId：job-ae9b-4302a6f8d819 提交时间：2017-01-06 15:17:35 开始时间：2017-01-06 15:17:40 结束时间：2017-01-06 15:17:52 任务耗时：12s 所属用户： 上卷状态：ROLLUP_OK	select userid, sum(charge) as charge from [illegible] where event_day=20170104 and query = '鲜花' group by userid order by charge desc limit 10	成功

查询结果集共10条数据，如下表所示

userid	charge
10622836	[illegible]
6383265	[illegible]
19156793	[illegible]
20456519	[illegible]
21748400	[illegible]
22143278	[illegible]
21242185	[illegible]



- 在‘userid’列上建立b+树索引
- 针对不同的列的数据特征，可以选择不同的索引类型

- 相较于原生Spark SQL获得了5倍左右的性能提升
- 百度3天的收费统计日志，共4TB数据，70000+文件 4TB，执行时间10~15秒

上卷表

700+ Columns

date	userid	searchid	baiduid	cmatch	...	shows	clicks	charge
1	1	1	10	2		10	1	5
1	1	2	11	3		10	1	5
1	1	3	12	2		10	1	5
1	1	4	13	1		10	1	5
1	1	5	14	1		10	1	5
1	2	6	14	2		10	1	5
1	2	7	15	3		10	1	5
1	2	8	16	4		10	1	5
1	2	9	17	5		10	1	5

Select date,userid,shows,clicks,charge from...

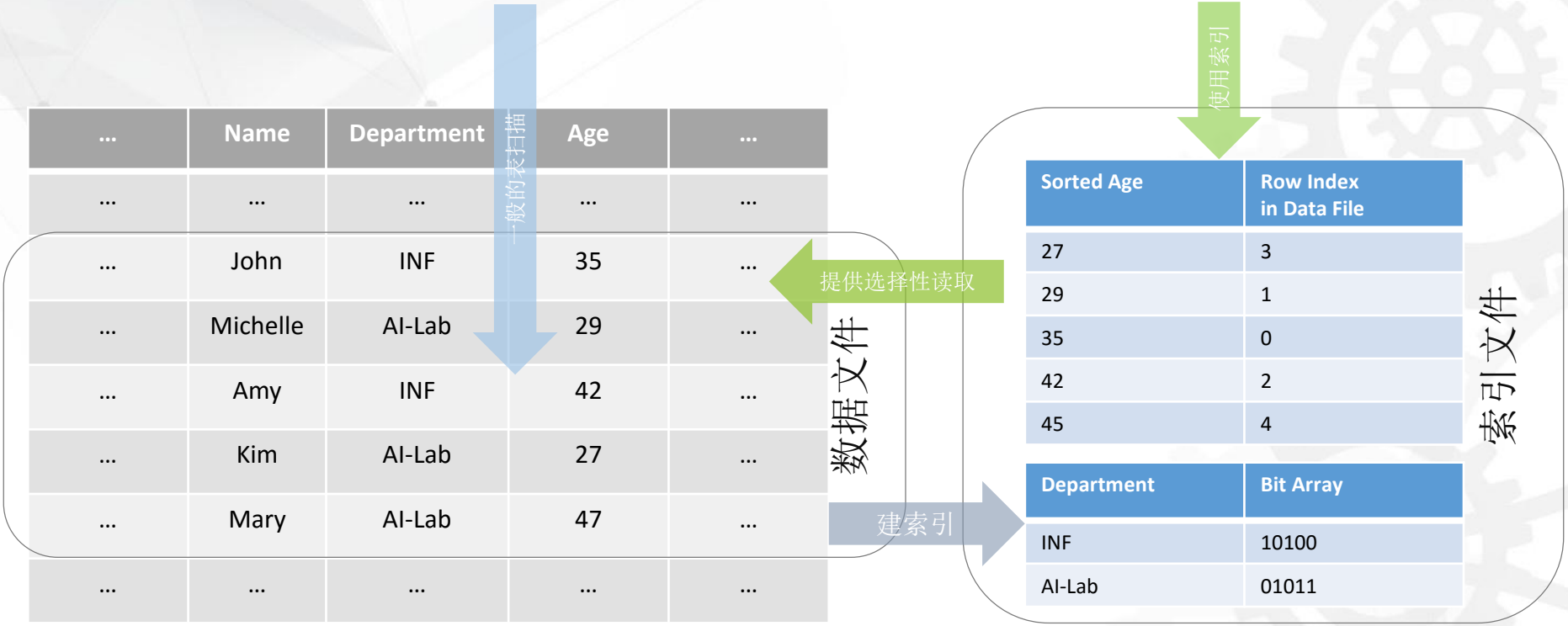
99% query 仅使用不到10列的数据

多维上卷表
(用户透明)

date	userid	shows	clicks	charge
1	1	50	5	25
1	2	40	4	20

date	cmatch	shows	clicks	charge
1	1	20	2	10
1	2	30	3	15
1	3	20	2	10
1	4	10	1	5
1	5	10	1	5

BigSQL中的Spinach



...	Name	Department	Age	...
...	...	...	...	...
...	John	INF	35	...
...	Michelle	AI-Lab	29	...
...	Amy	INF	42	...
...	Kim	AI-Lab	27	...
...	Mary	AI-Lab	47	...
...	...	...	...	...

Sorted Age	Row Index in Data File
27	3
29	1
35	0
42	2
45	4

Department	Bit Array
INF	10100
AI-Lab	01011

Select xxx from xxx where age > 29 and department in (INF, AI-Lab)

BigSQL中的Spinach

...	Name	Department	Age	...
...	...	...	...	...
...	John	INF	35	...
...	Michelle	AI-Lab	29	...
...	Amy	INF	42	...
...	Kim	AI-Lab	27	...
...	Mary	AI-Lab	47	...
...	...	...	...	...

数据文件

读取到缓存

内存缓存

Department	Row Index in Data File
INF	2
AI-Lab	3

Age	Row Index in Data File
35	0
29	1

Agenda

- 背景介绍
- Spinach介绍
- Spark和Spinach在百度
  - Spark在百度
  - 百度BigSQL
  - BigSQL中的优化
- 未来计划

未来计划

- 兼容更多数据文件格式
- 提供分层缓存机制与缓存管理
- 基于索引提供针对常见SQL算子的物理计划优化
- 与流式处理的融合

THANKS

SequeMedia 盛拓传媒 IT168.com ITPUB ChinaUnix