SPARK SUMMIT 2017

OAP: Optimized Analytics Package for Spark Platform

Daoyuan Wang (Intel)
Yuanjian Li (Baidu)

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.

About me

Daoyuan Wang

- developer@Intel
- Focuses on Spark optimization
- An active Spark contributor since 2014

Yuanjian Li

- Baidu INF distributed computation
- Apache Spark contributor
- Baidu Spark team leader

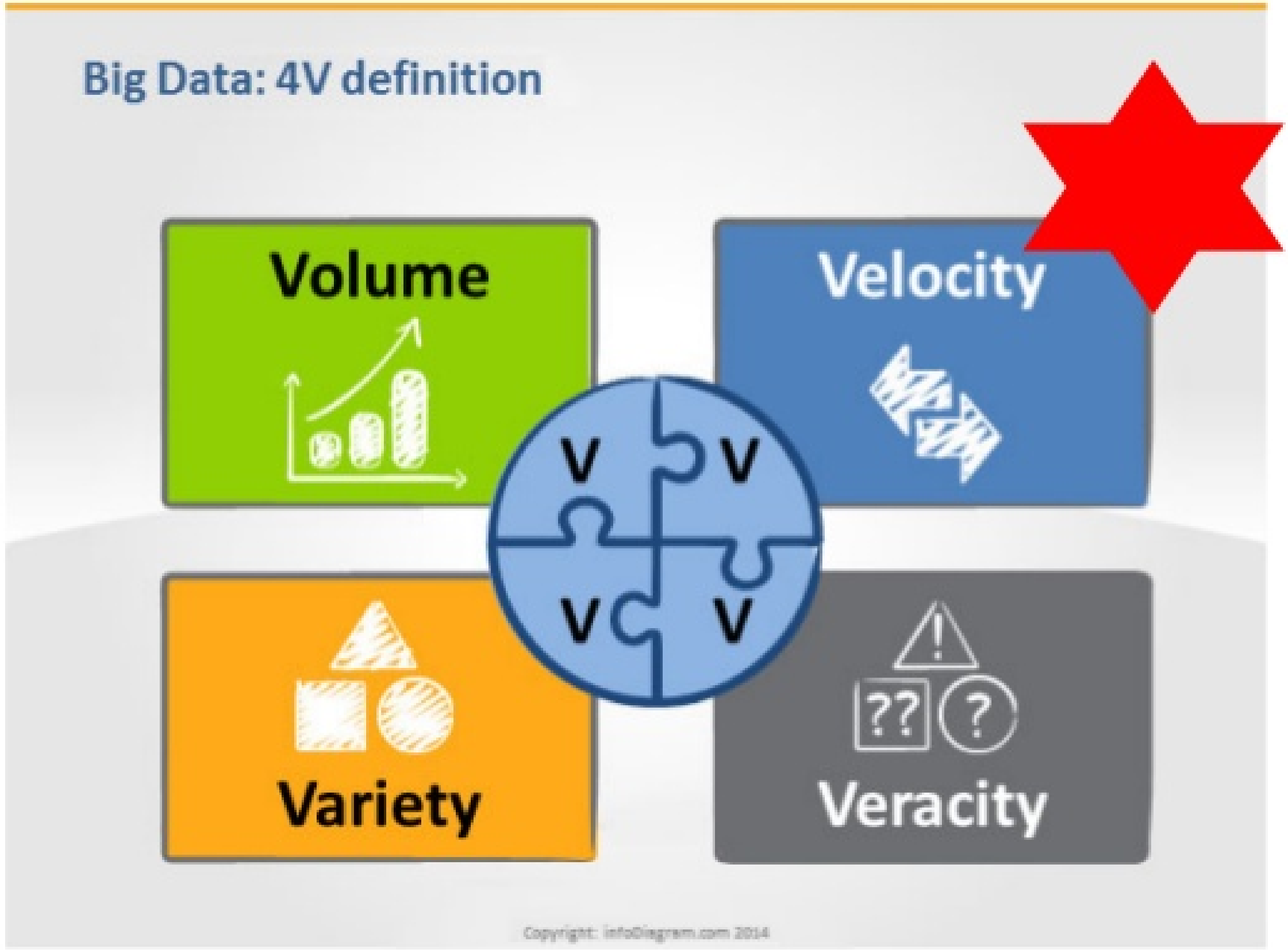
Agenda

- Background for OAP
- Key features
- Benchmark
- OAP and Spark in Baidu
- Future plans

Agenda

- Background for OAP
- Key features
- Benchmark
- OAP and Spark in Baidu
- Future plans

Data Analytics in Big Data Definition



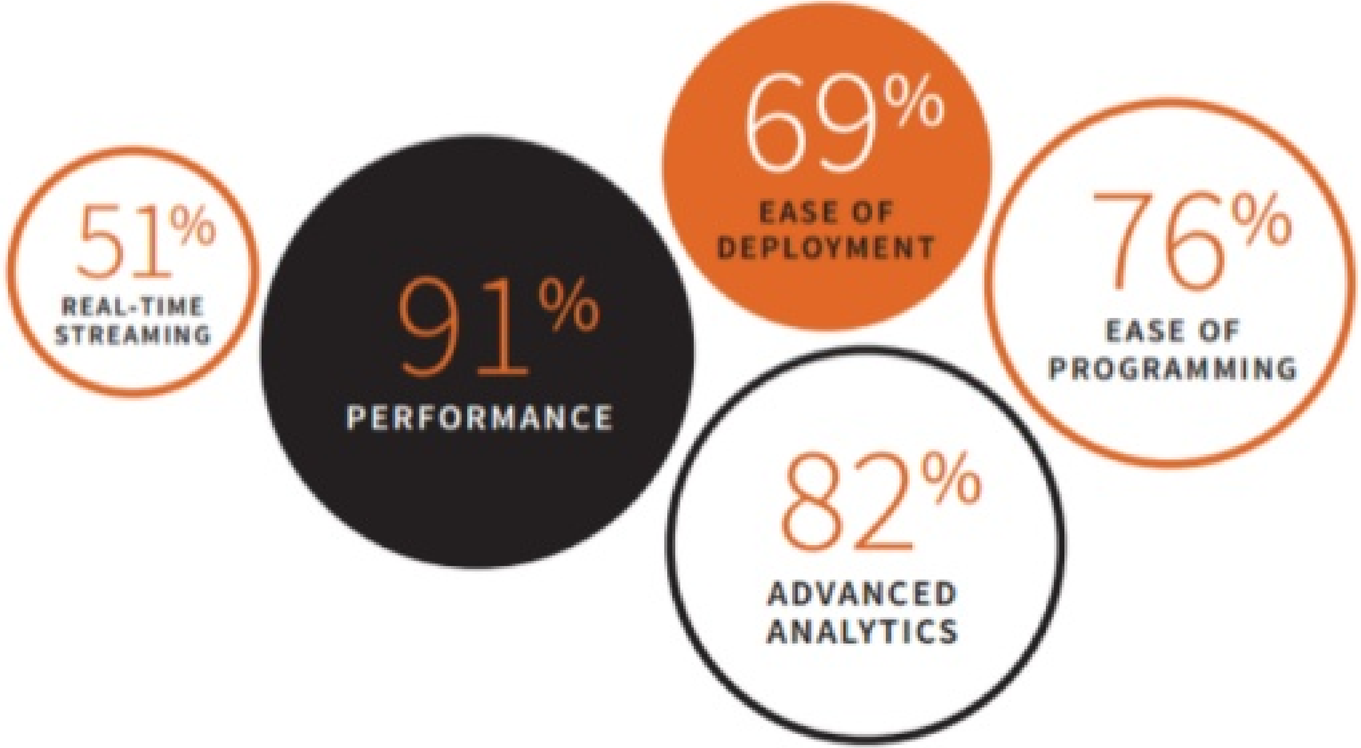
Copyright: infoDiagram.com 2014

- People wants OLAP against large dataset **as fast as possible**.
- People wants extract information from new coming data **as soon as possible**.

Data Analytics Acceleration is Required by Spark Users

FEATURES USERS CONSIDER IMPORTANT

Respondents were allowed to select more than one feature.



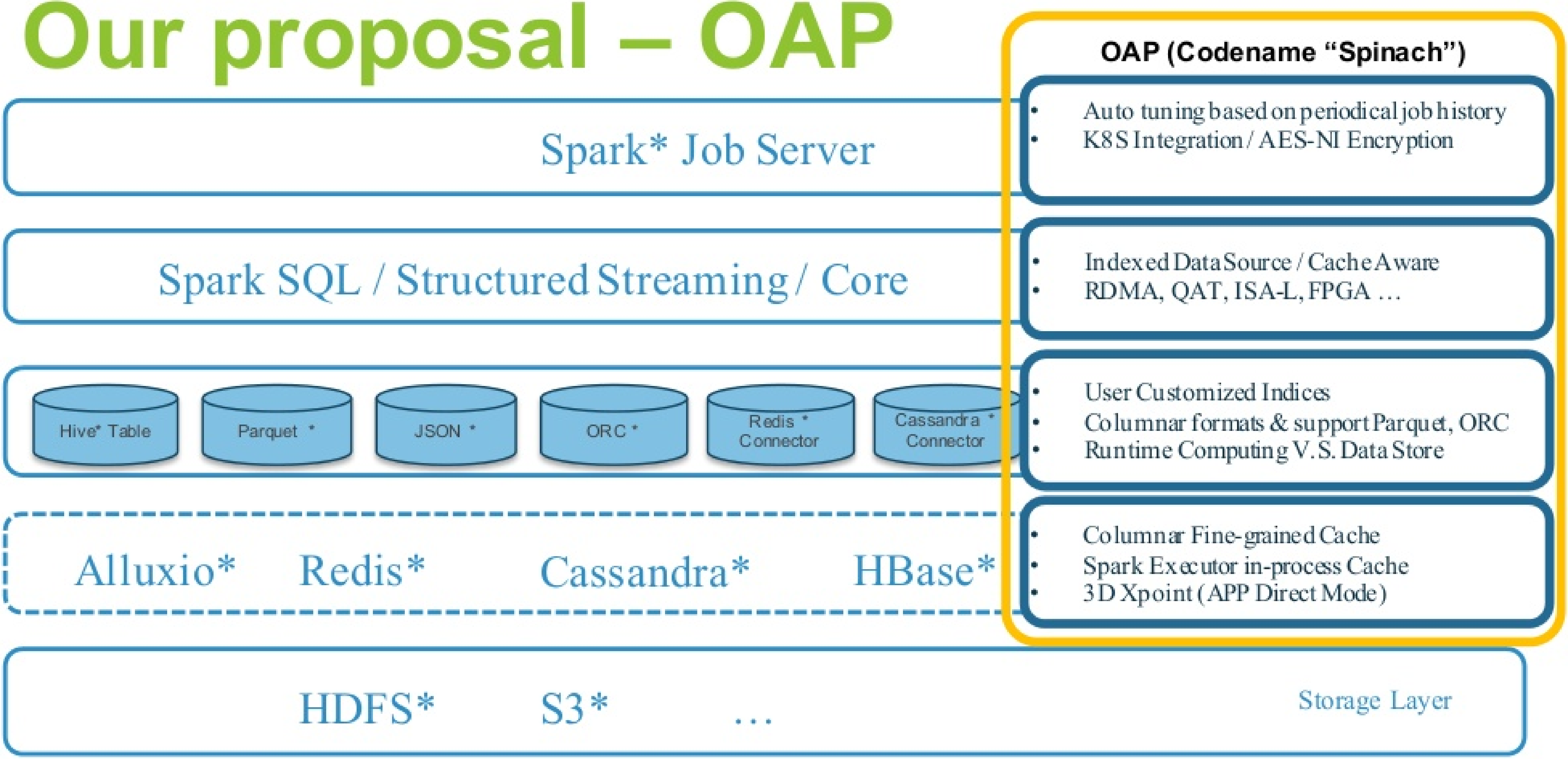
http://cdn2.hubspot.net/hubfs/438089/DataBricks_Surveys_-_Content/2016_Spark_Survey/2016_Spark_Infographic.pdf

Emerging hardware technology

Intel® Optane™ Technology
Data Center Solutions

Accelerate applications for fast caching and storage, reduce transaction costs for latency-sensitive workloads and increase scale per server. Intel® Optane™ technology allows data centers to deploy bigger and more affordable datasets to gain new insights from large memory pools.

Our proposal – OAP



Spark* Job Server

Spark SQL / Structured Streaming / Core

Hive* Table

Parquet *

JSON *

ORC *

Redis * Connector

Cassandra * Connector

Alluxio* Redis* Cassandra* HBase*

HDFS* S3* …

Storage Layer

OAP (Codename "Spinach")

- Auto tuning based on periodical job history
- K8S Integration / AES-NI Encryption

- Indexed DataSource / Cache Aware
- RDMA, QAT, ISA-L, FPGA …

- User Customized Indices
- Columnar formats & support Parquet, ORC
- Runtime Computing V.S. Data Store

- Columnar Fine-grained Cache
- Spark Executor in-process Cache
- 3D Xpoint (APP Direct Mode)

Why OAP

Low cost

- Makes full use of existing hardware
- Open source

Good Performance

- Index just like traditional database
- Up to 5x boost in real-world

Easy to Use

- Easy to deploy
- Easy to maintain
- Easy to learn

Agenda

- Background for OAP
- Key features
- Benchmark
- OAP and Spark in Baidu
- Future plans

A Simple Example

1. Run with OAP

```
    $SPARK_HOME/sbin/start-thriftserver --package oap.jar;
```

2. Create a OAP table

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

3. Create a single column B+ Tree index

```
    beeline> CREATE SINDEX idx_1 ON src (a) USING BTREE;
```

4. Insert data

```
    beeline> INSERT INTO TABLE src SELECT key, value FROM xxx;
```

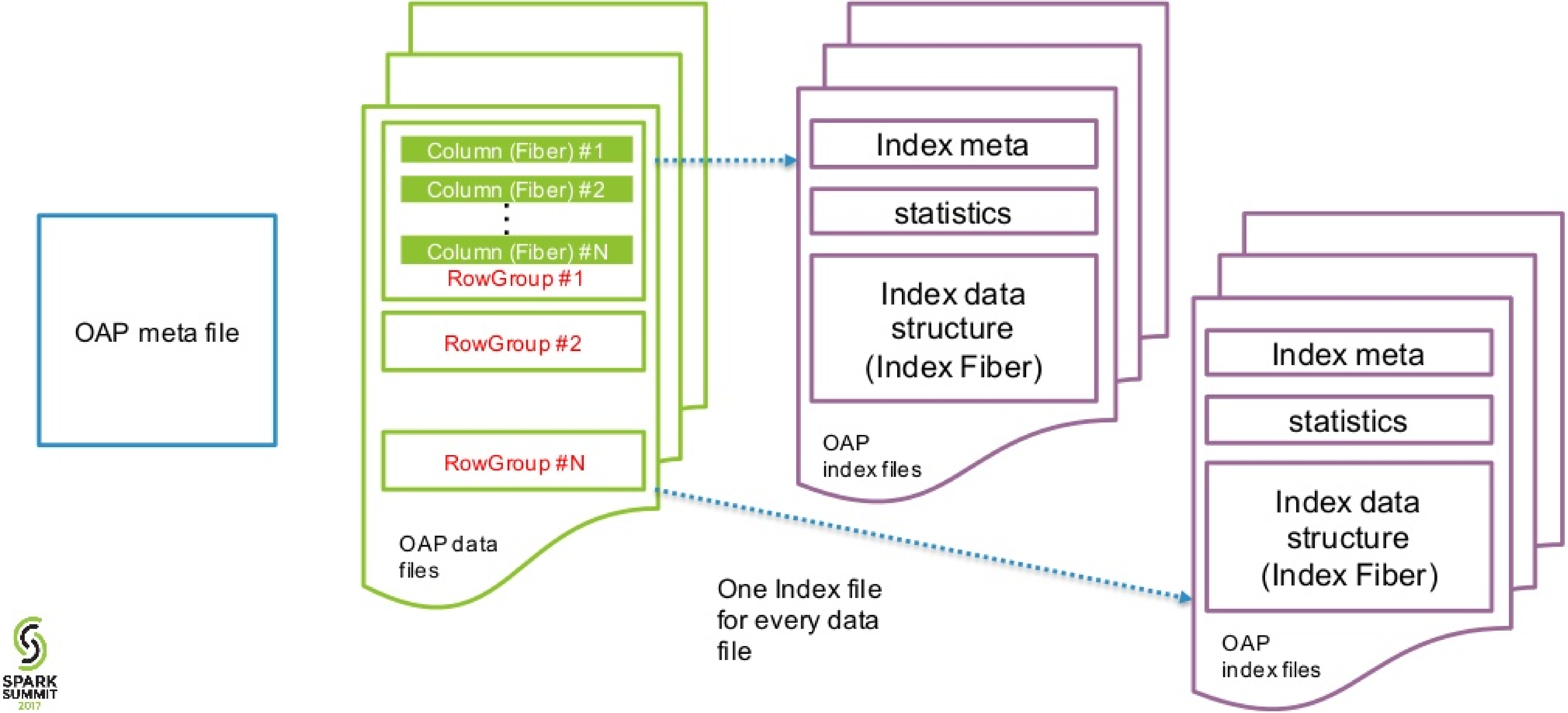
5. Refresh index

```
    beeline> REFRESH SINDEX on src;
```

6. Execution would automatically utilize index

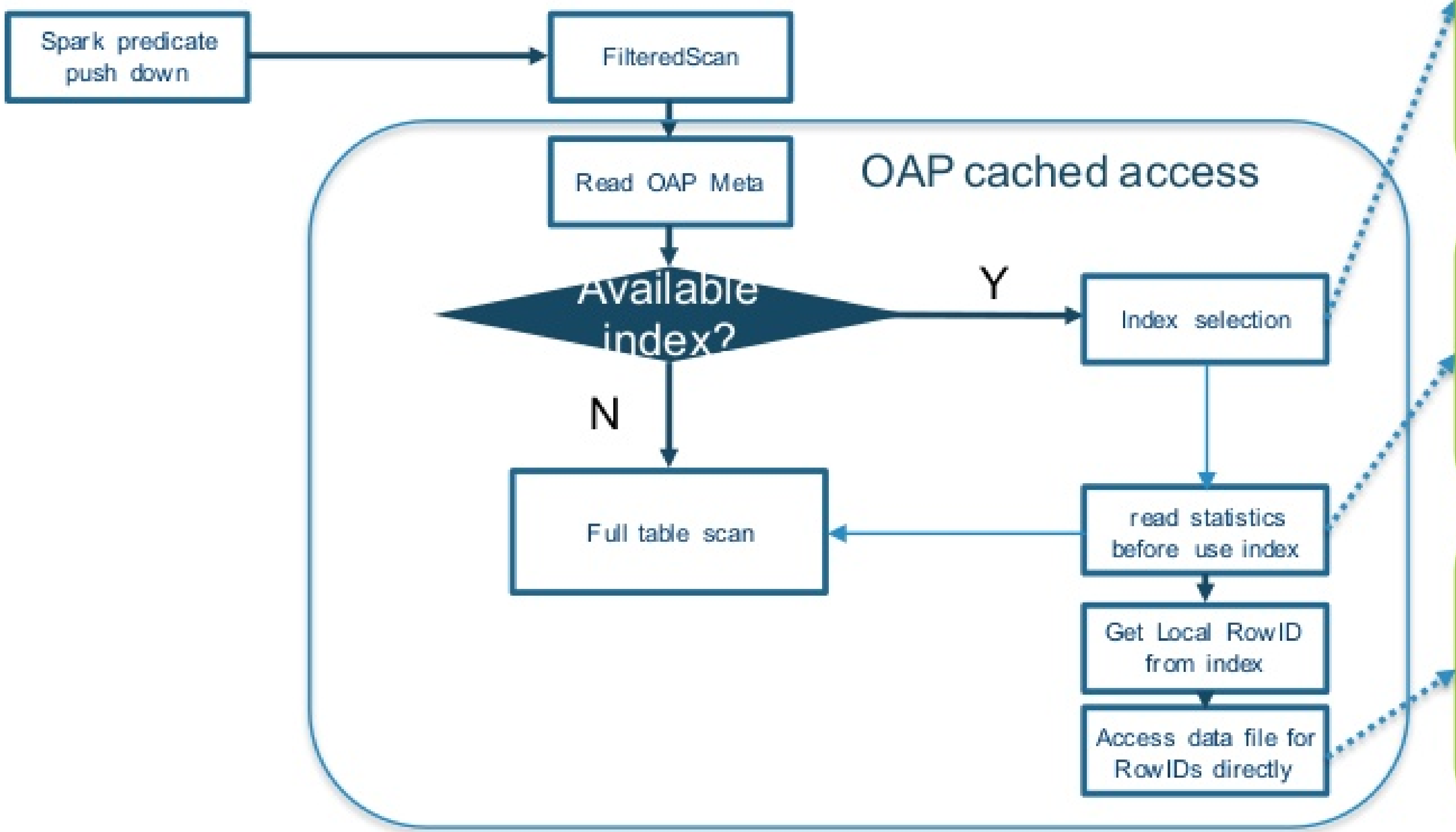
```
    beeline> SELECT MAX(value), MIN(value) FROM src WHERE a > 100 and a <
1000;
```

OAP Files and Fibers



One Index file for every data file

OAP Internals - index

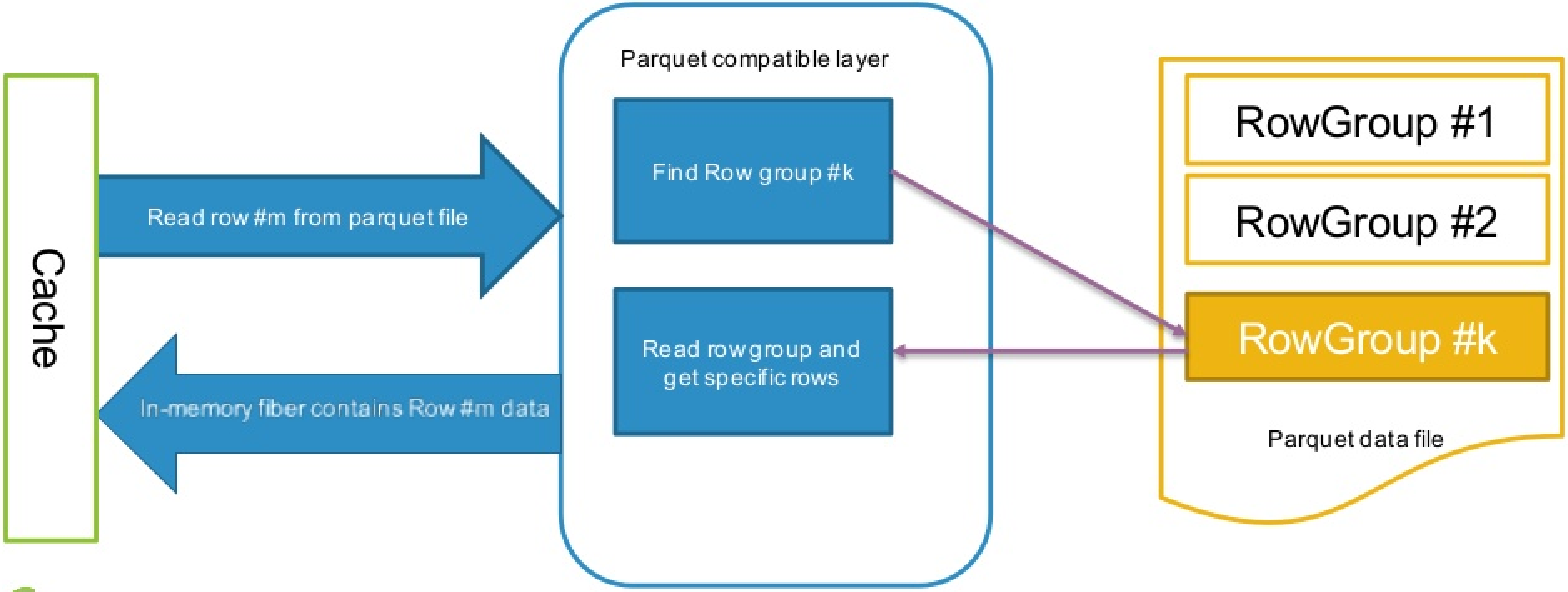

Spark predicate push down
FilteredScan
OAP cached access
Read OAP Meta
Available index?
Y
N
Index selection
Full table scan
read statistics before use index
Get Local RowID from index
Access data file for RowIDs directly

Supports Btree Index and BitMap Index, find best match among all created indices

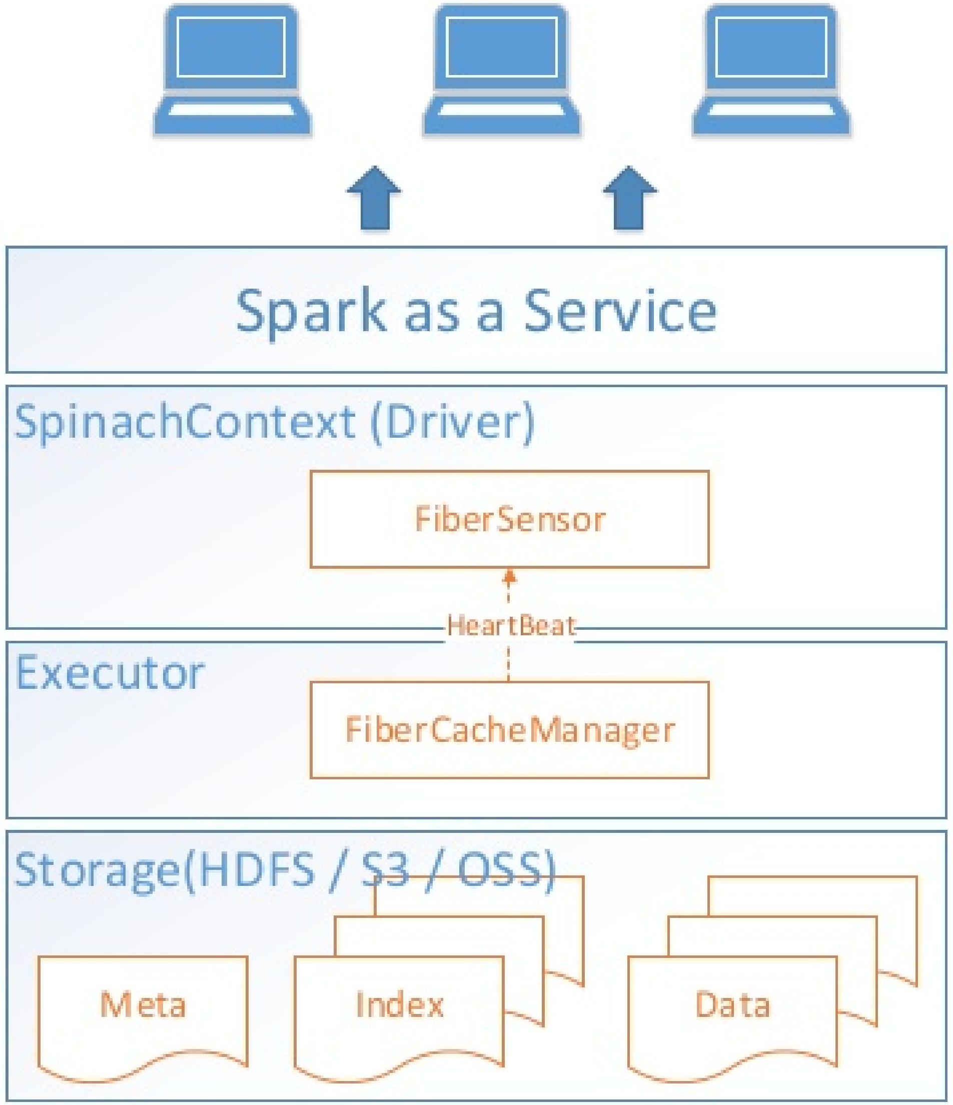
Supports statistics such as MinMax, PartbyValue, Sample, BloomFilter

Only reads data fibers we need and puts those fibers into cache (in-memory fiber)

OAP compatible layer


Cache
Read row #m from parquet file
In-memory fiber contains Row #m data
Parquet compatible layer
Find Row group #k
Read rowgroup and get specific rows
RowGroup #1
RowGroup #2
RowGroup #k
Parquet data file

OAP Data locality



Agenda

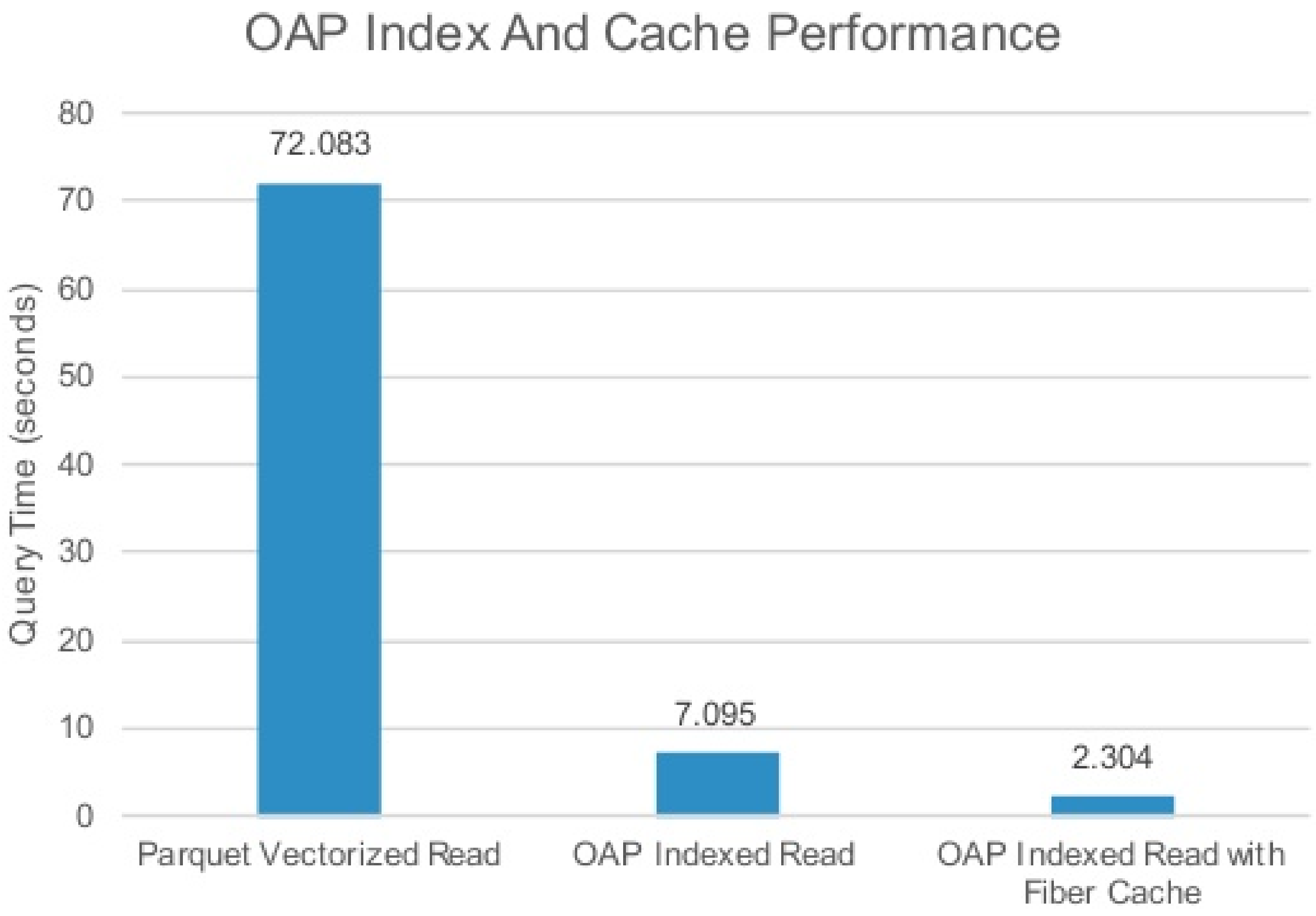
- Background for OAP
- Key features
- Benchmark
- OAP and Spark in Baidu
- Future plans

Performance

Cluster:
1 Master + 2 Slaves

Hardware:
CPU – 2x E5-2699 v4
RAM – 256 GB
Storage – S3610 1.6TB

Data:
300GB (Compressed Parquet)
2 Billion Records

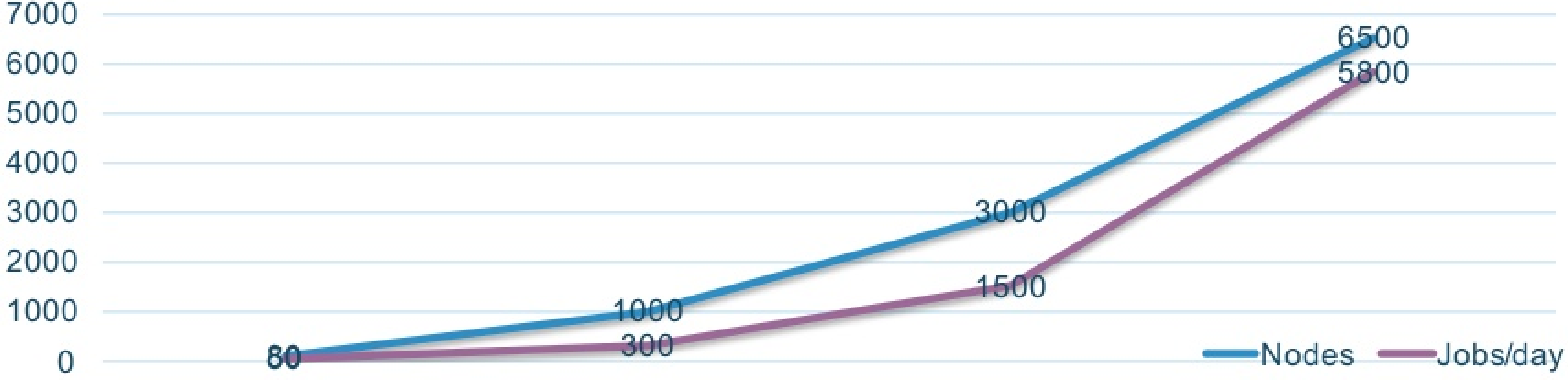


SPARK SUMMIT 2017

Agenda

- Background for OAP
- Key features
- Benchmark
- OAP and Spark in Baidu
- Future plans

Spark In Baidu



2014

- Spark import to Baidu
- Version: 0.8

2015

- Build standalone cluster
- Integrate with in-house FS\Pub-Sub\DW
- Version: 1.4

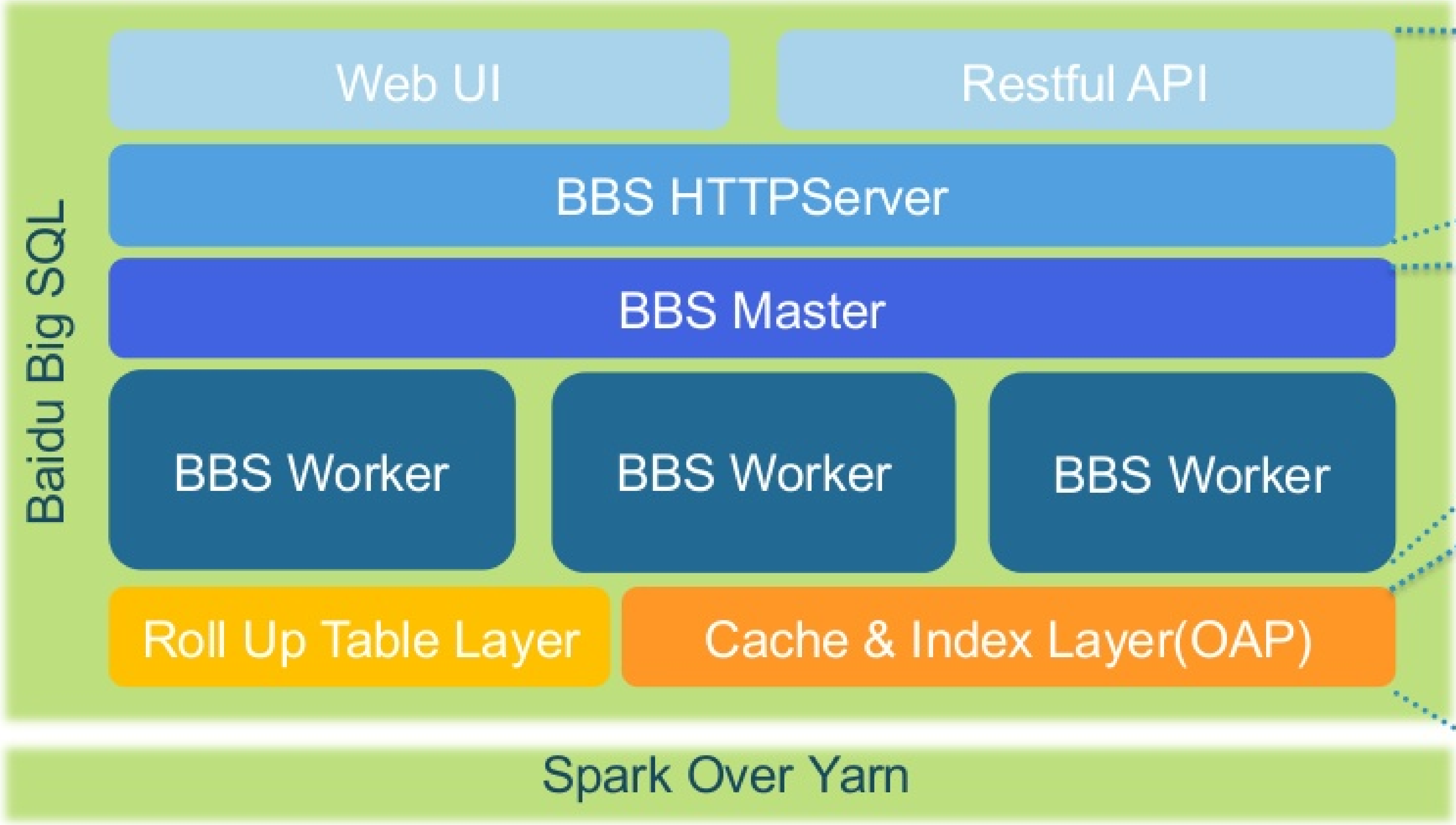
2016

- Build Cluster over YARN
- Integrate with in-house Resource Scheduler System
- Version: 1.6

2017

- SQL\Graph Service over Spark
- OAP
- Version: 2.1

Baidu Big SQL



API Layer:

- Meta Control API
- Job API:

Load\Export\Query\Index Control

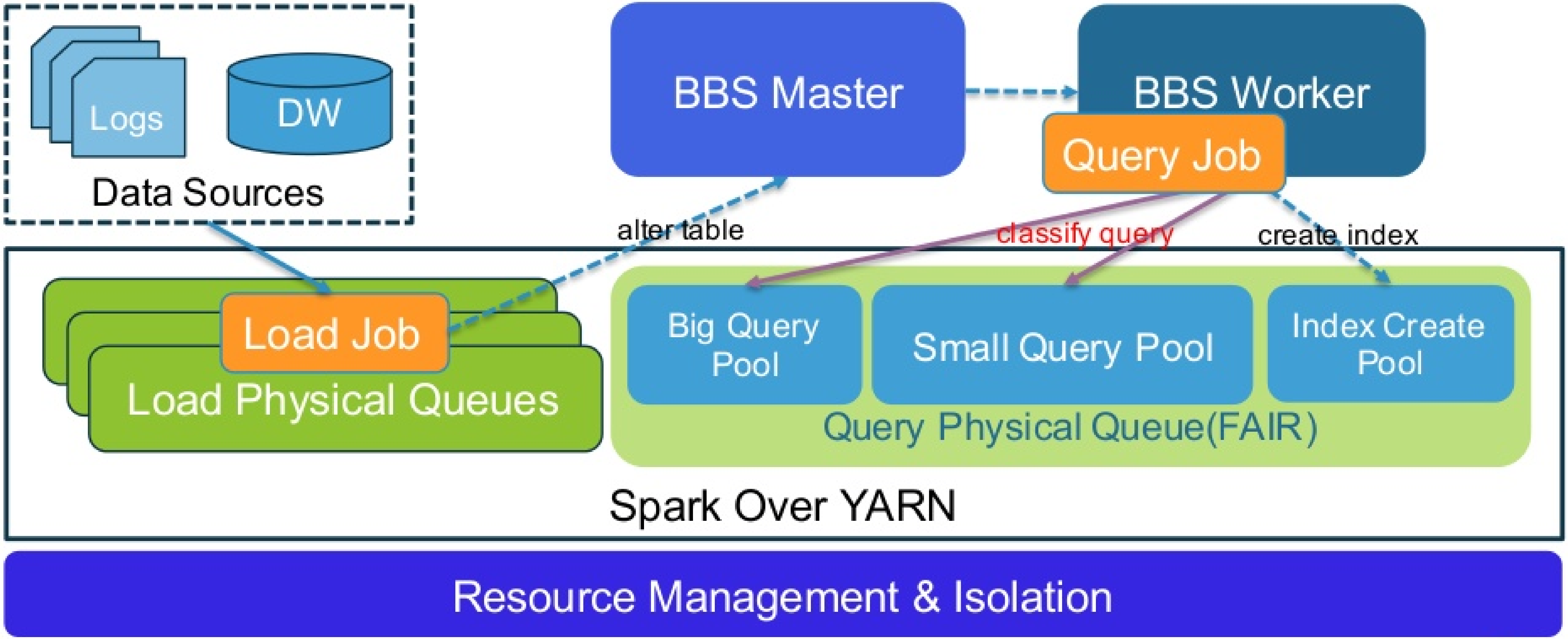
Control Layer:

- Meta Control
- Job Scheduler
- Spark Driver
- Query Classification

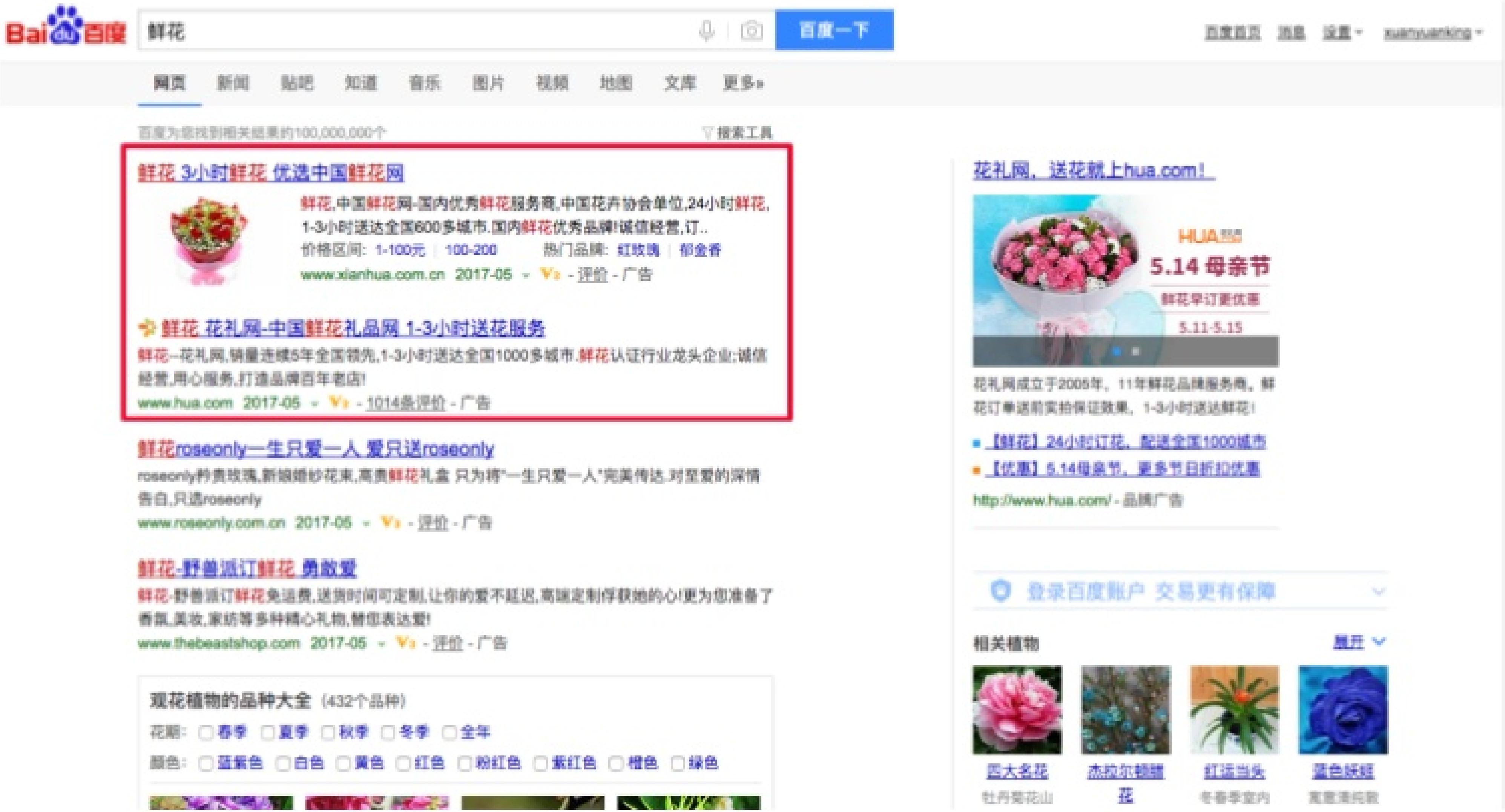
Boosting Layer:

- Roll Up Table Management
- Roll Up Query Change
- Index Create\Update
- Cache Hit

Baidu Big SQL



Introductory Story



百度为您找到相关结果约100,000,000个

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 - V2 - 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 - V3 - 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,只送roseonly

www.roseonly.com.cn 2017-05 - V3 - 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 - V3 - 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！

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

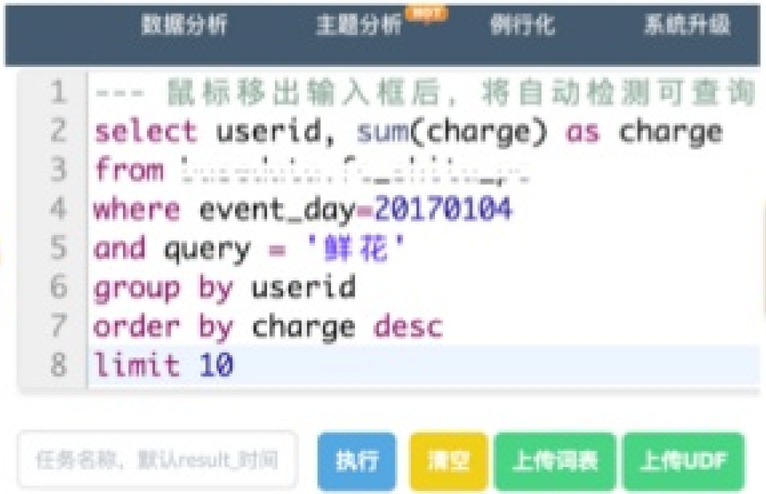
登录百度账户 交易更有保障

相关植物 展开

四大名花	杰拉尔顿蜡花	红运当头	蓝色妖姬
牡丹菊花山		冬春季室内	寓意清纯脱

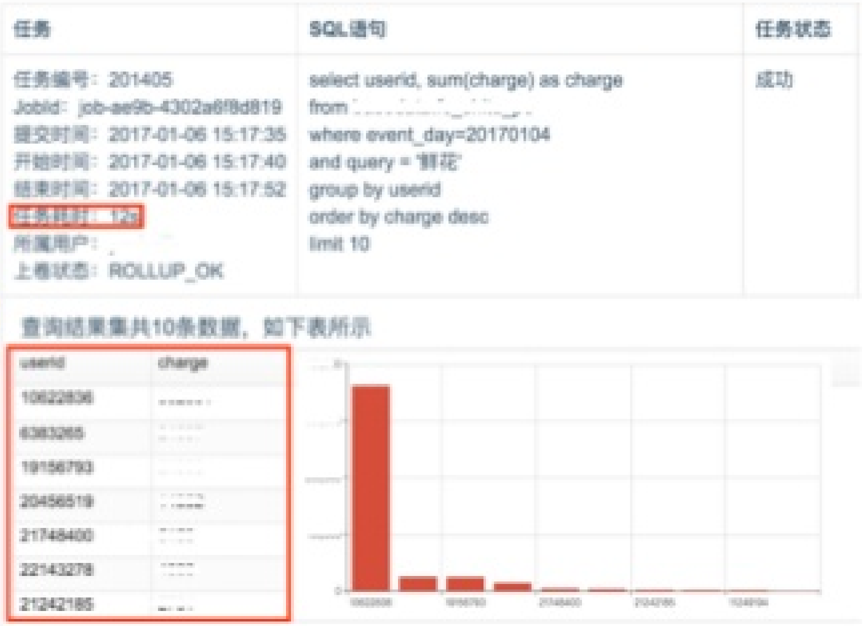
Introductory Story

Get the top 10 charge sum and correspond advertiser which triggered by the query word ‘flower’



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

- Create index on ‘userid’ column
- Various index types to choose for different fields types



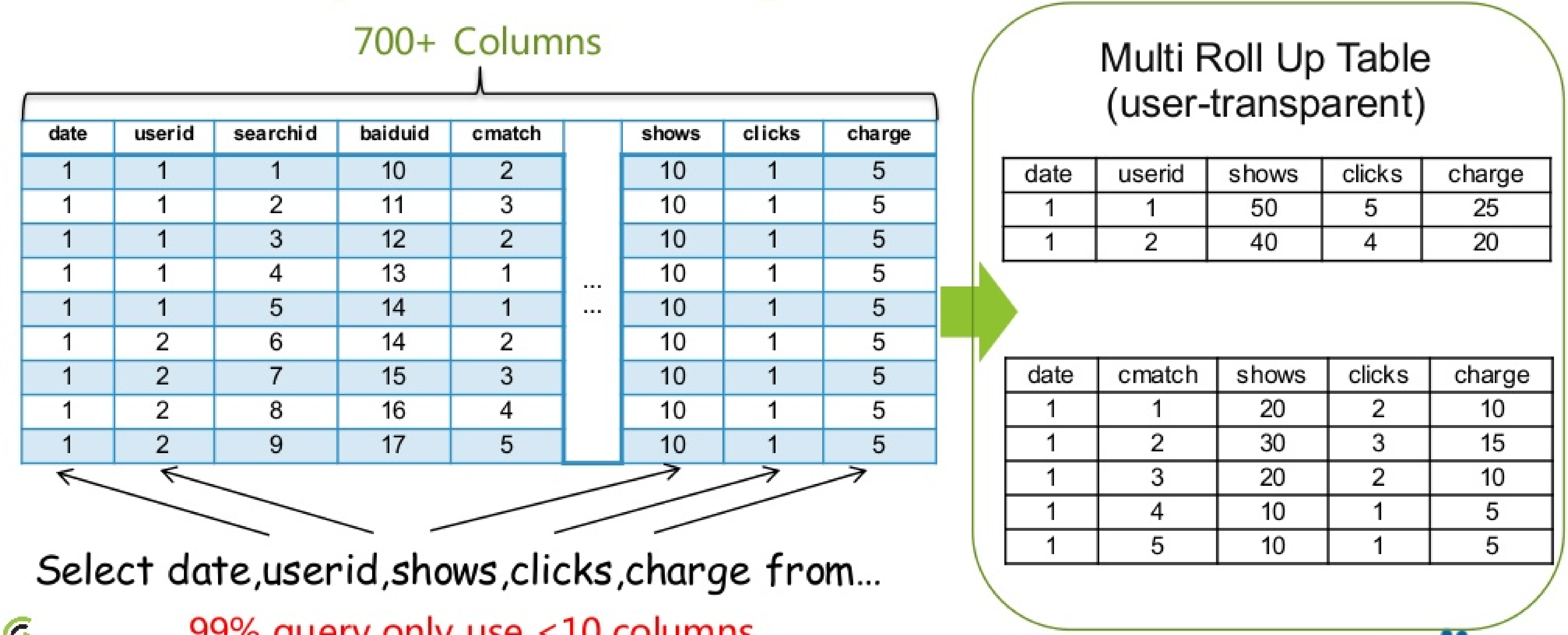
任务	SQL语句	任务状态
任务编号：201405 JobId：job-ae9b-4302a6f8d819 提交时间：2017-01-06 15:17:35 开始时间：2017-01-06 15:17:40 结束时间：2017-01-06 15:17:52 任务耗时：12s 所属用户：[redacted] 上卷状态：ROLLUP_OK	select userid, sum(charge) as charge from [redacted] where event_day=20170104 and query = '鲜花' group by userid order by charge desc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

查询结果集共10条数据，如下表所示

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

- ×5 speed boosting than native spark sql, ×80 than MR Job
- 3 day baidu charging log, 4TB data,70000+ files, query time in 10~15s

Roll Up Table Layer



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…

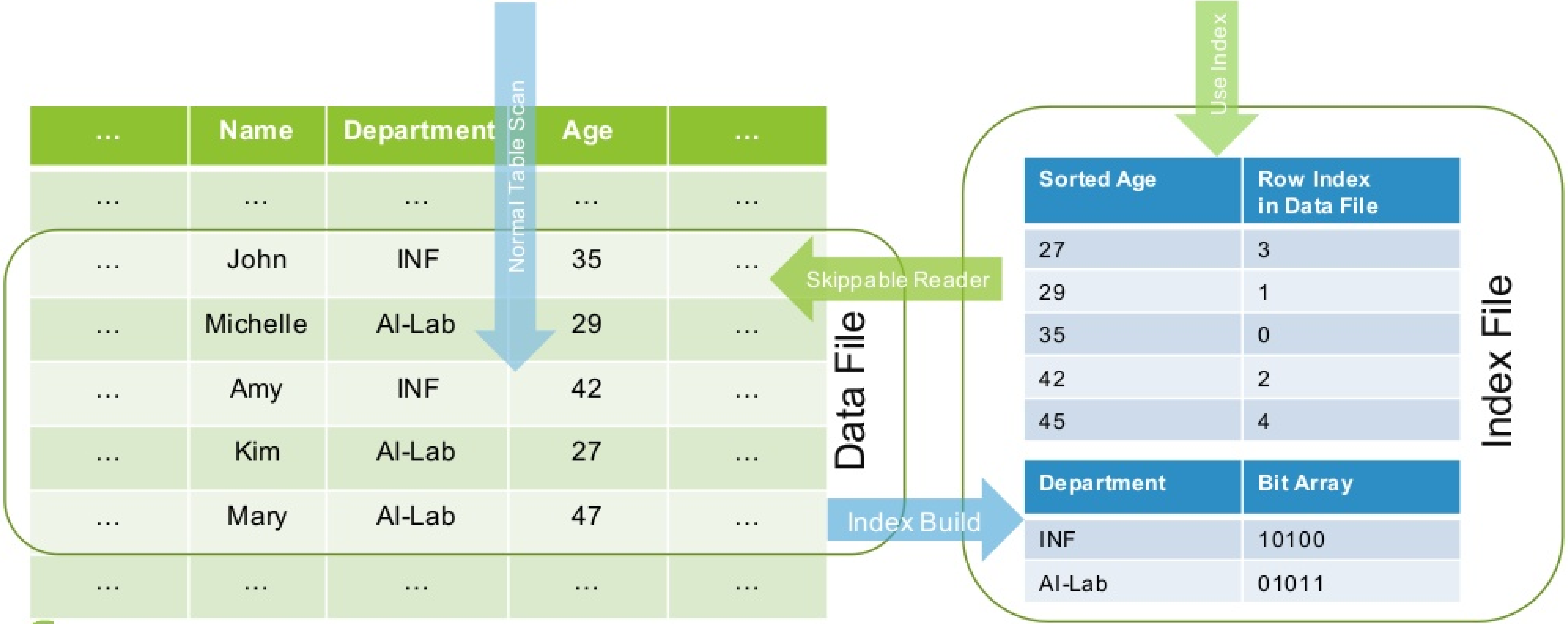
Multi Roll Up Table
(user-transparent)

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

99% query only use <10 columns

OAP In BigSQL



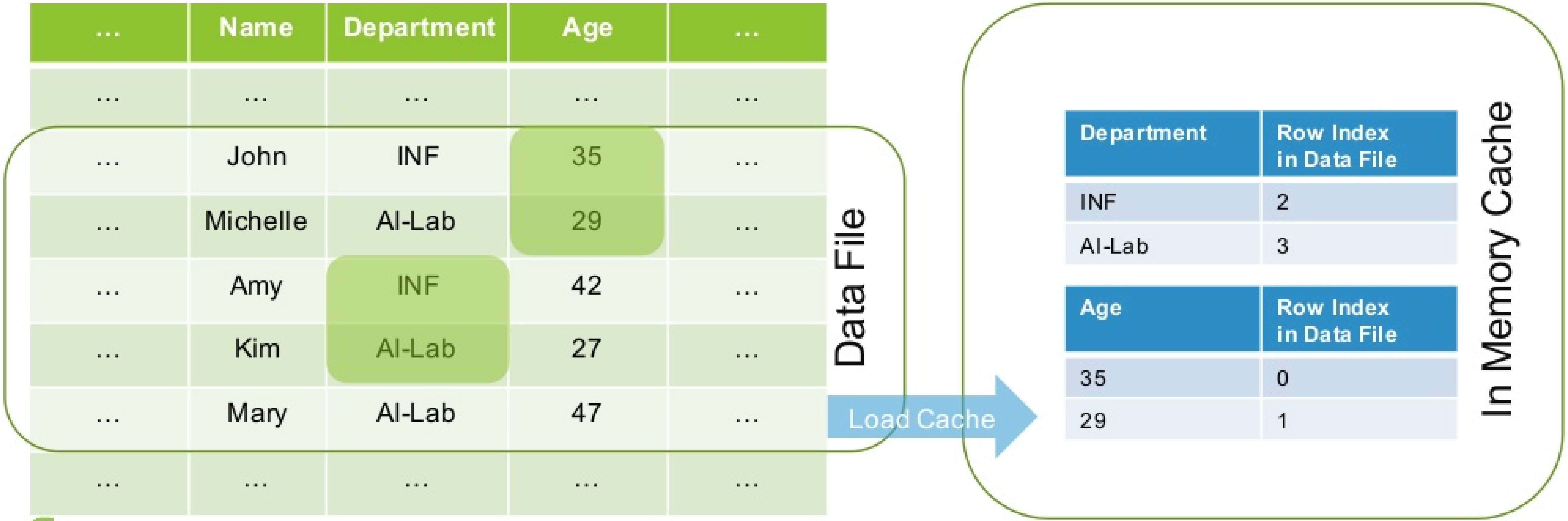
…	Name	Department	Age	…
…	…	…	…	…
…	John	INF	35	…
…	Michelle	AI-Lab	29	…
…	Amy	INF	42	…
…	Kim	AI-Lab	27	…
…	Mary	AI-Lab	47	…
…	…	…	…	…

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)

OAP In BigSQL



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

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

Age	Row Index in Data File
35	0
29	1

BBS’s Contribute to Spark

- Spark-4502

Spark SQL reads unneccesary nested fields from Parquet

- Spark-18700

getCached in HiveMetastoreCatalog not thread safe cause driver OOM

- Spark-20408

Get glob path in parallel to reduce resolve relation time

- …

Agenda

- Background for OAP
- Key features
- Benchmark
- OAP and Spark in Baidu
- Future plans

Future plans

- Compatible with more data formats
- Explicit cache and cache management
- Optimize SQL operators (join, aggregate) with index
- Integrate with structured streaming
- Utilize Latest hardware technology, such as Intel QAT or 3D XPoint.
- Welcome to contribute!

https://github.com/Intel-bigdata/OAP

WOMEN IN BIG DATA

Connect
Inspire
Build
Cultivate
Create
Grow
Embrace
Educate
Lead

WOMEN IN BIG DATA NETWORKING LUNCHEON

The Women in Big Data team invites you to join us for lunch, network with your peers and hear from a dynamic panel of experts. Come learn what career & growth opportunities are available in the field of big data analytics.

Agenda:

12.20PM	Grab Lunch & Networking
12:30PM-12:40PM	Women in Big Data Overview with Soumya Guptha, Marketing Manager, Intel
12:40PM-12:45PM	My journey in Data Analytics & Artificial Intelligence with Ziya Ma, Intel VP & Director, Big Data Technologies
12:50PM-01:40PM	Panel: Making The Best Out Of The Fast Paced Data World!

Panel: Making The Best Out Of The Fast Paced Data World!

Gayle Sheppard, VP, New Technology Group, Intel | Ritika Gunnar, Global VP of IBM Cloud and Cognitive, IBM | Eva Tse, Director of Big Data Services, Netflix | Jennifer Shin, CEO 8 path solutions | Soumya Guptha, Marketing Manager, Software and Solutions Group, Intel

Join us for a networking luncheon to hear from industry experts from leading companies such as IBM, Intel and others on their investments in Big Data technologies such as Spark, Machine Learning, Artificial Intelligence.

www.womeninbigdata.org/ | @DataWomen | Women in Big Data Forum | www.meetup.com/Women-in-Big-Data-Meetup/

GRASSROOTS COMMUNITY CHAMPIONING WOMEN'S LEADERSHIP AND SUCCESS IN BIG DATA

SPARK SUMMIT 2017

Thank You.

daoyuan.wang@intel.com

liyuanjian@baidu.com