

Apache Kylin v2.x

加速大数据OLAP分析

李栋 | Dong Li

技术合伙人 & 高级架构师

Apache Kylin是什么？

OSC 源创会
Opensource Innovation Meetup

IT大咖说
知识分享平台

BI
可视化
(SQL)

Interactive

Reporting

Dashboard

OLAP引擎

Apache Kylin

Hadoop

HDFS

Hive

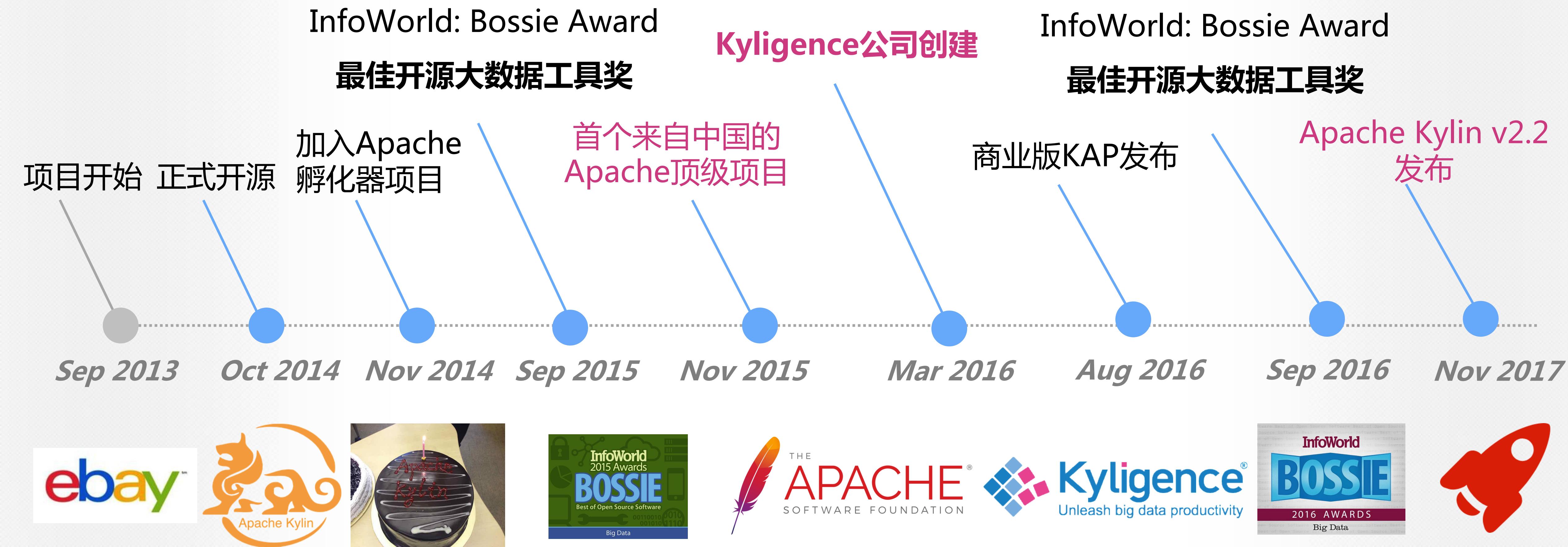
HBase

- 3 万亿条数据,
< 1 秒 查询延迟
@头条, 国内第一新闻资讯app
- 60+ 维度的Cube
@CPIC
- JDBC / ODBC / RestAPI
- BI 集成
Tableau、Excel、Cognos、Superset

Apache Kylin 历史

OSC 源创会
Opensource Innovation Meetup

IT大咖说
知识分享平台



Apache Kylin全球案例

osc 源创会
Opensource Innovation Meetup

IT大咖说
知识分享平台

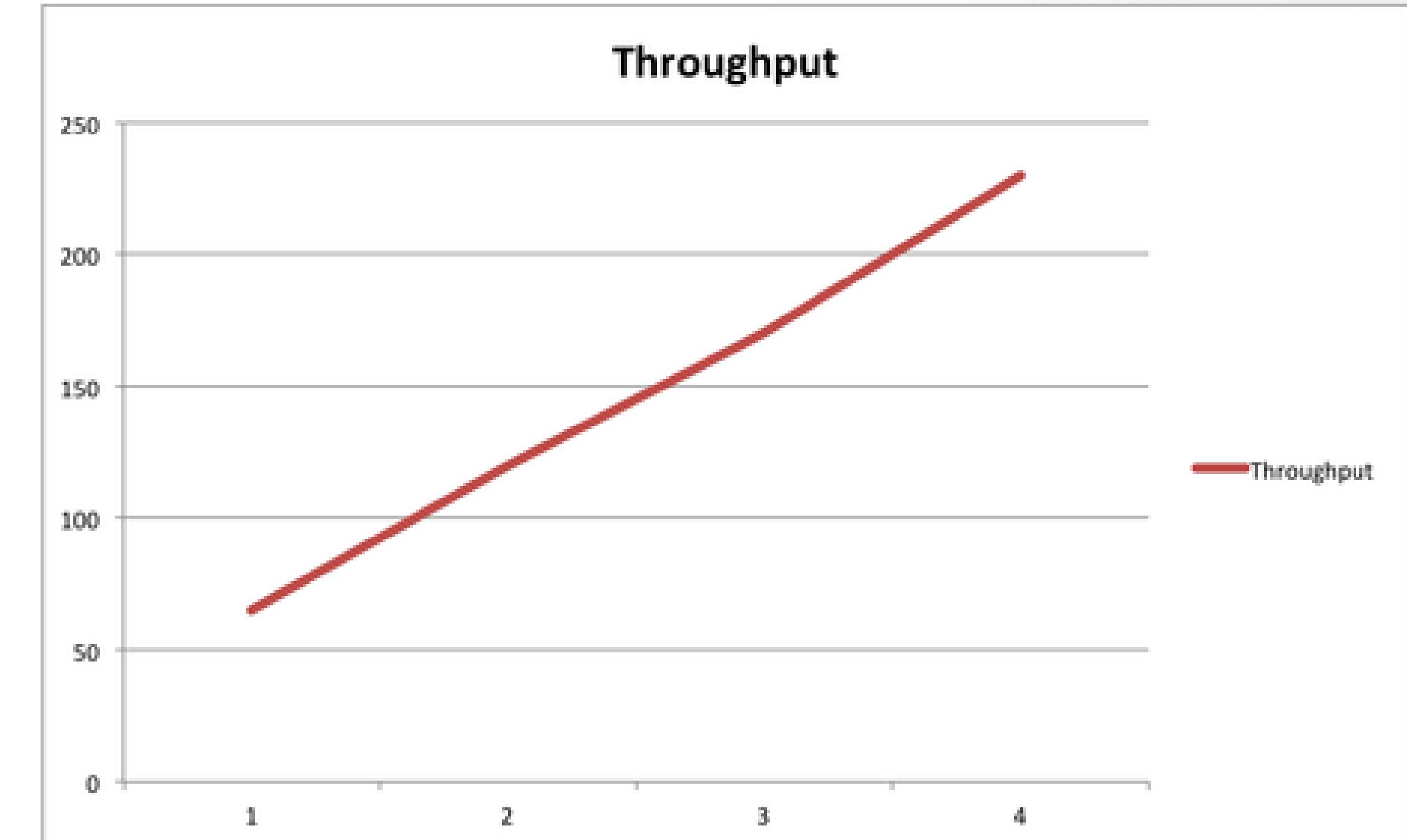
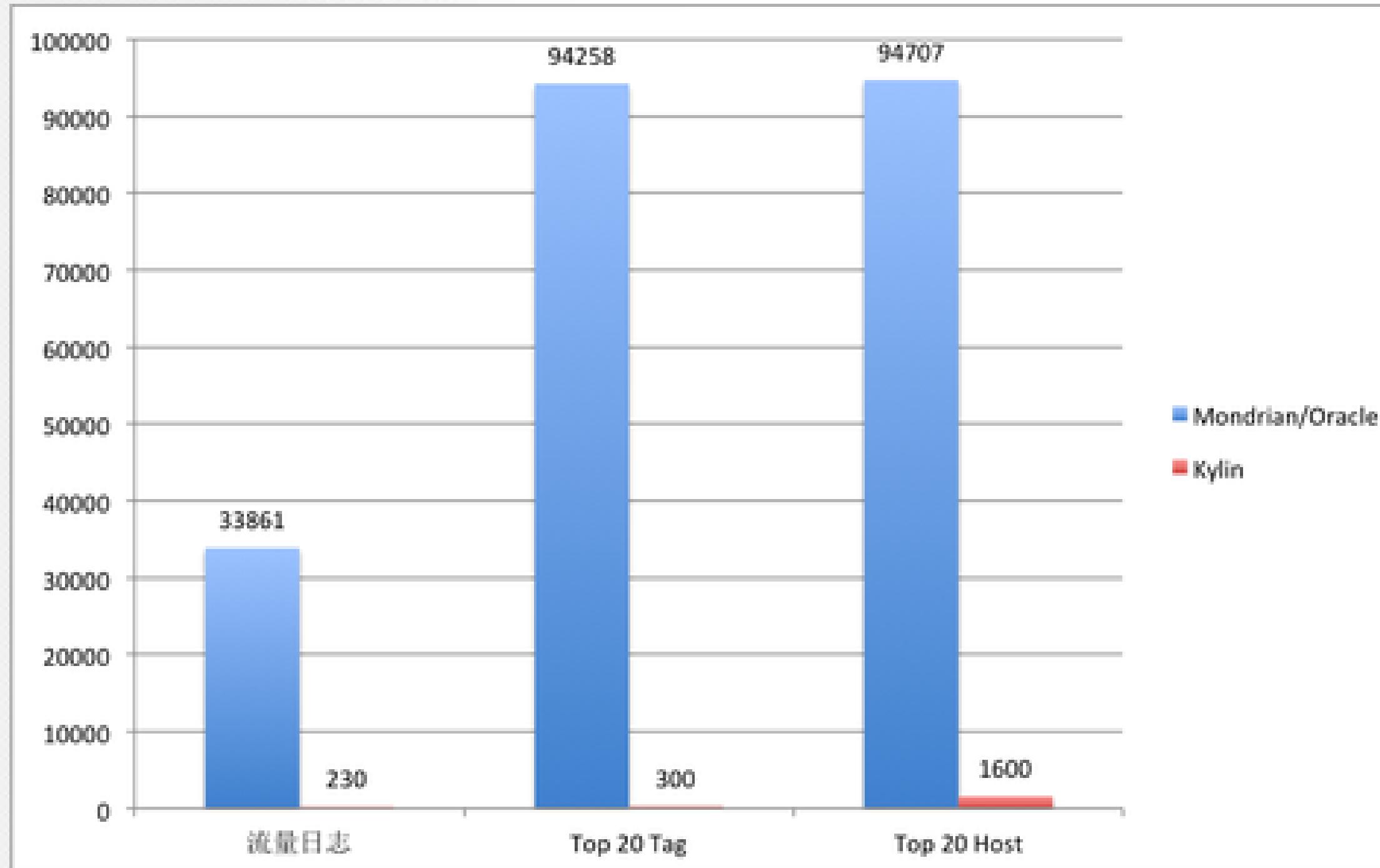


全球 500+ 用户部署到生产环境

Apache Kylin 性能测试

osc 源创会
Opensource Innovation Meetup

IT大咖说
知识分享平台

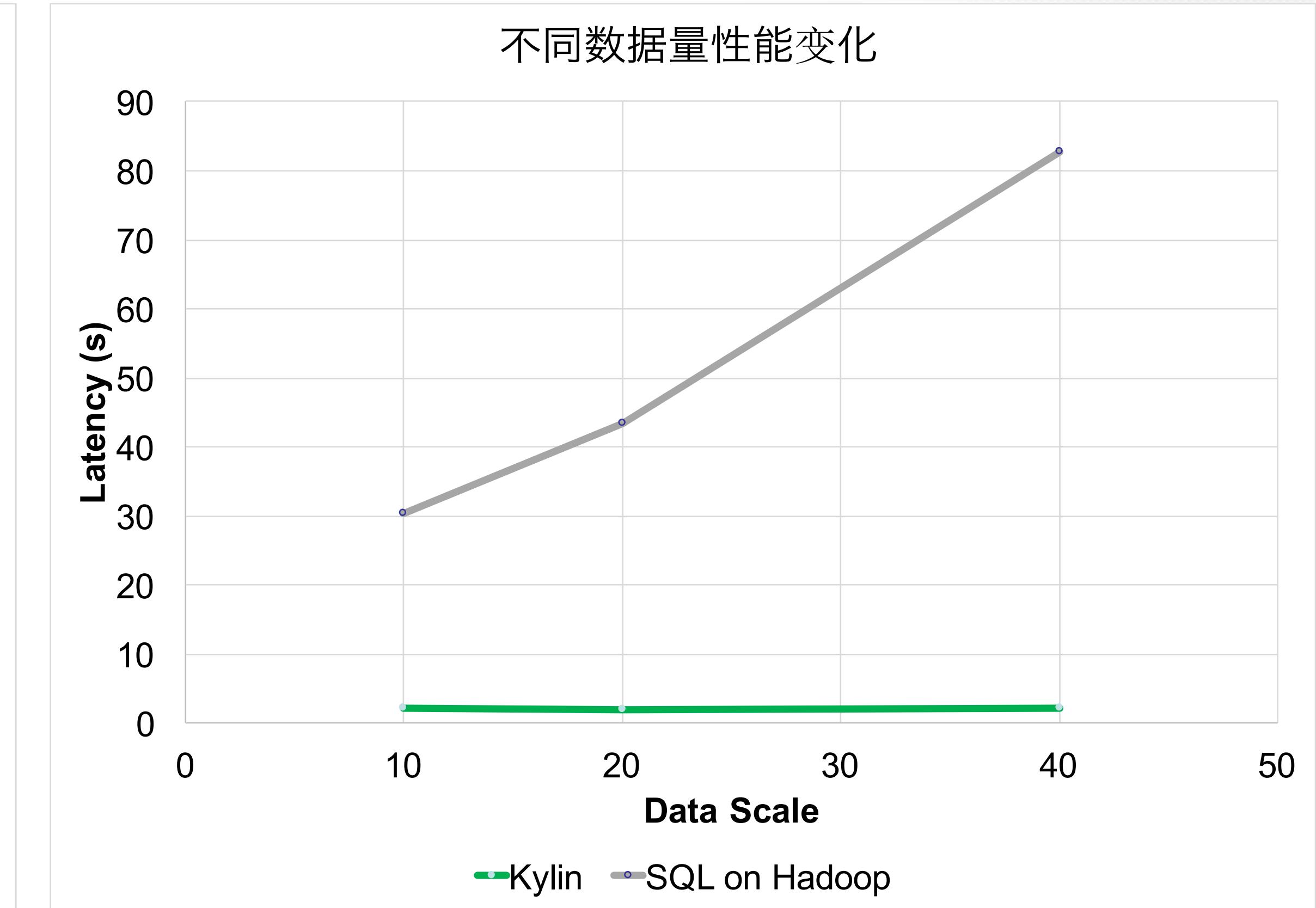
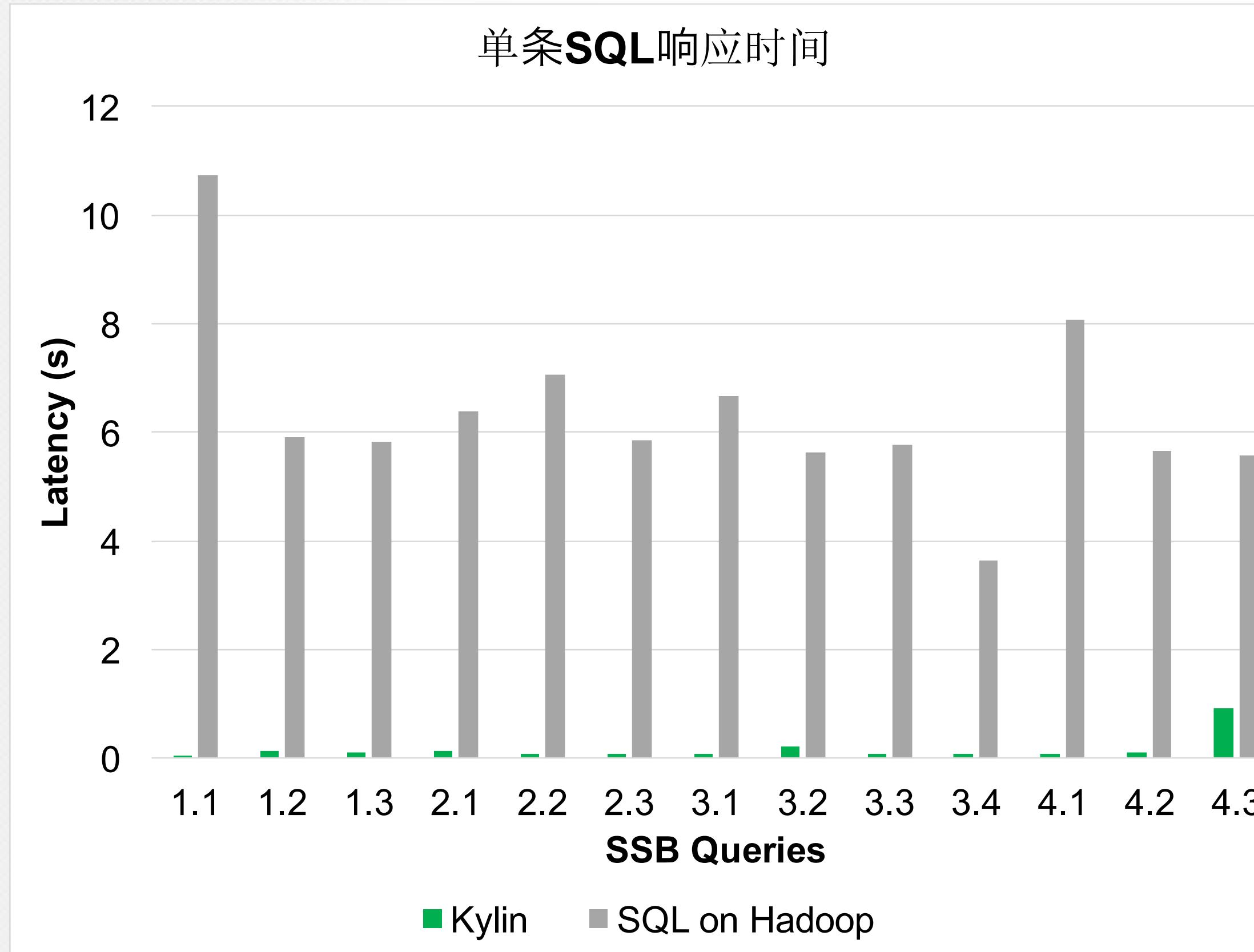


By 网易:
<http://datalab.int-yt.com/archives/1708>

Apache Kylin vs. SQL-on-Hadoop

源创会
Opensource Innovation Meetup

IT大咖说
知识分享平台



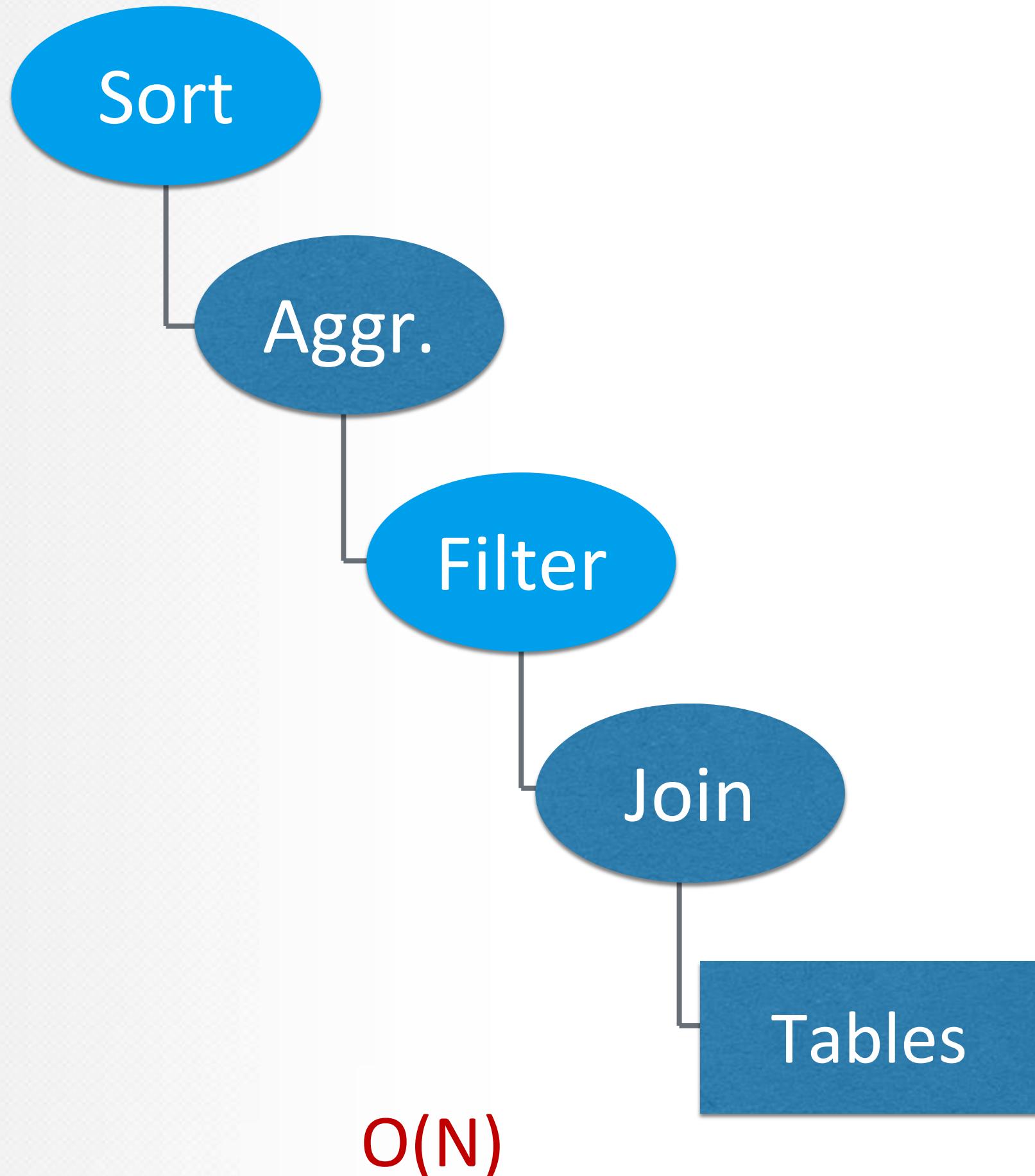
Star schema benchmark:

<http://www.cs.umb.edu/~poneil/StarSchemaB.PDF>

Apache Kylin 为什么快

OSC 源创会
Opensource Innovation Meetup

IT大咖说
知识分享平台



样例：

分析一段时间内，不同“*returnflag*”和“*orderstatus*”对应的销售情况

select

```
I_returnflag,  
o_orderstatus,  
sum(I_quantity) as sum_qty,  
sum(I_extendedprice) as sum_base_price  
...
```

from

```
v_lineitem  
inner join v_orders on I_orderkey = o_orderkey
```

where

```
I_shipdate <= '1998-09-16'
```

group by

```
I_returnflag,  
o_orderstatus
```

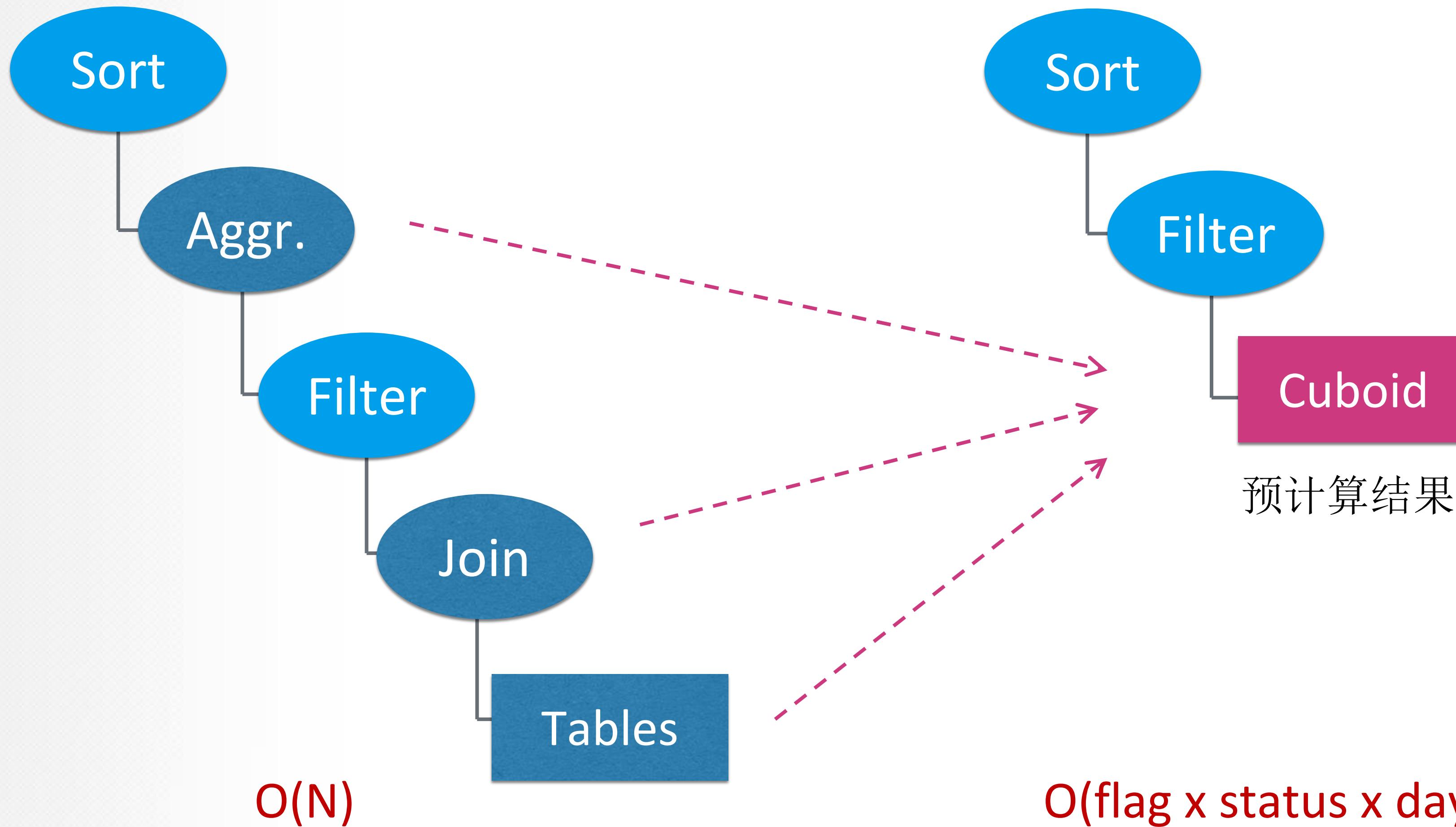
order by

```
I_returnflag,  
o_orderstatus;
```

Apache Kylin 为什么快

OSC 源创会
Opensource Innovation Meetup

IT 大咖说
知识分享平台

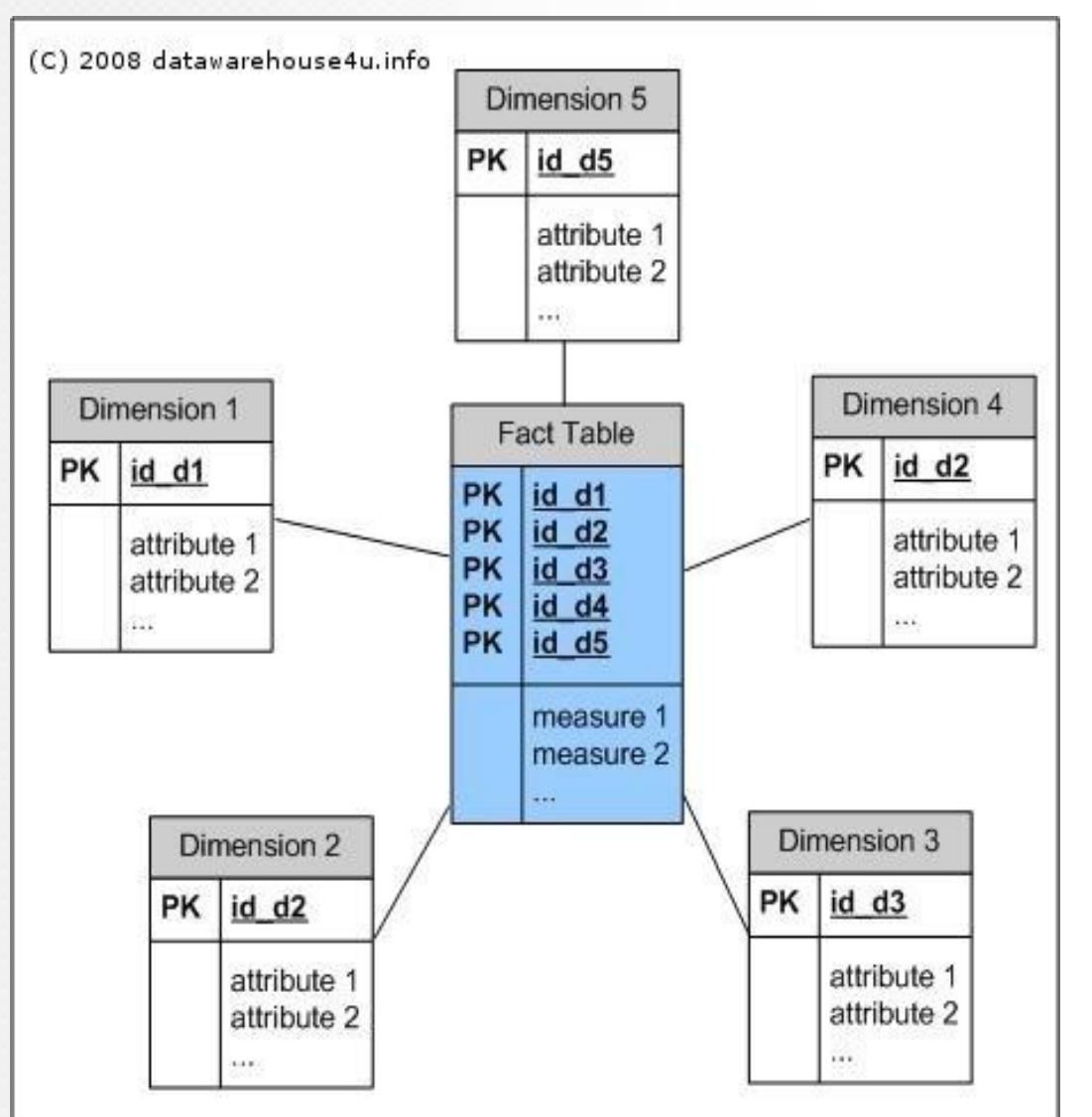


Apache Kylin 关键在于预计算

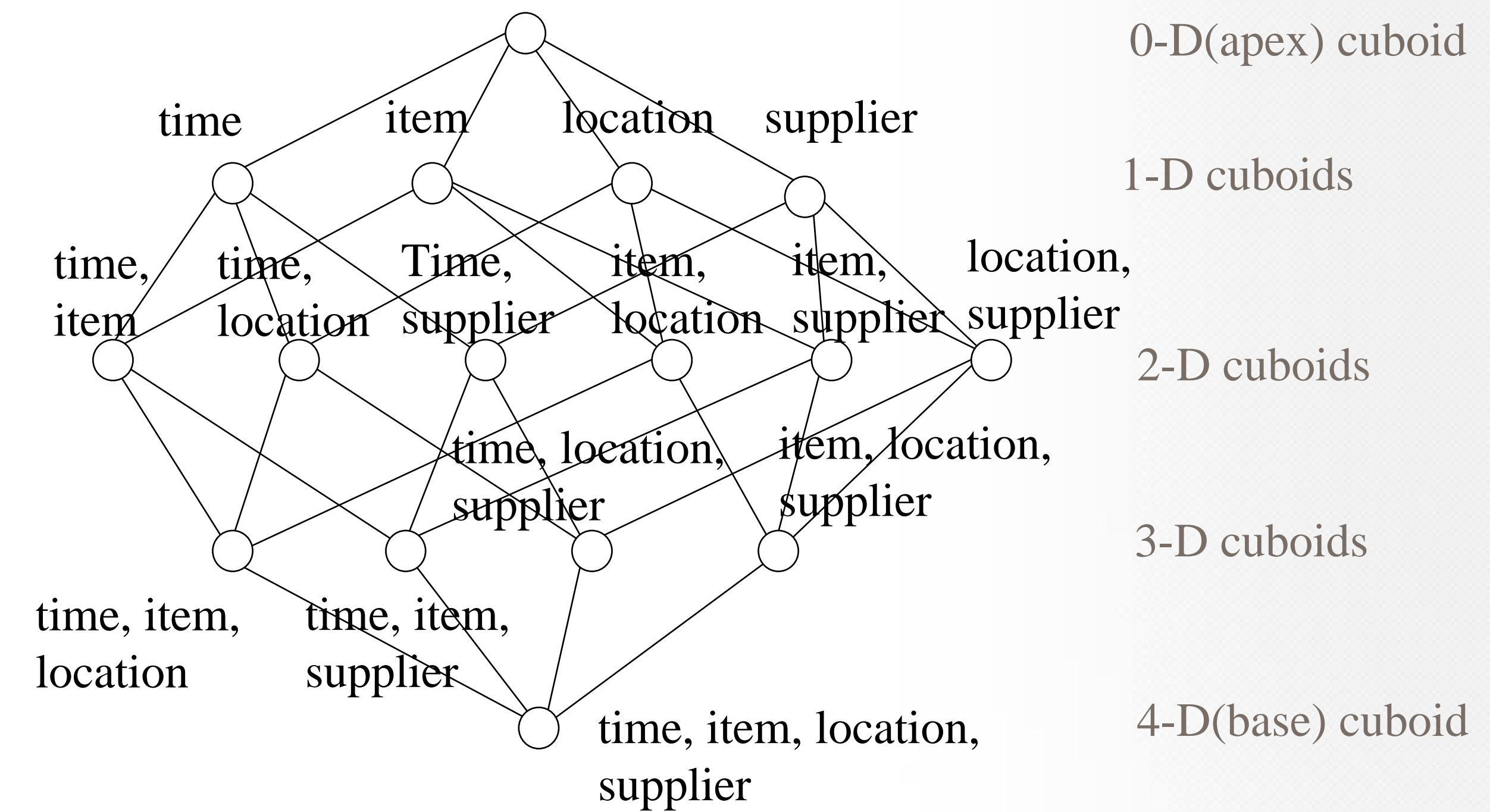
OSC 源创会
Opensource Innovation Meetup

IT 大咖说
知识分享平台

- OLAP Cube 理论基础
- Model 和 Cube 定义预计算范围
- Build Engine 执行预计算任务
- Query Engine 在预计算结果上完成查询



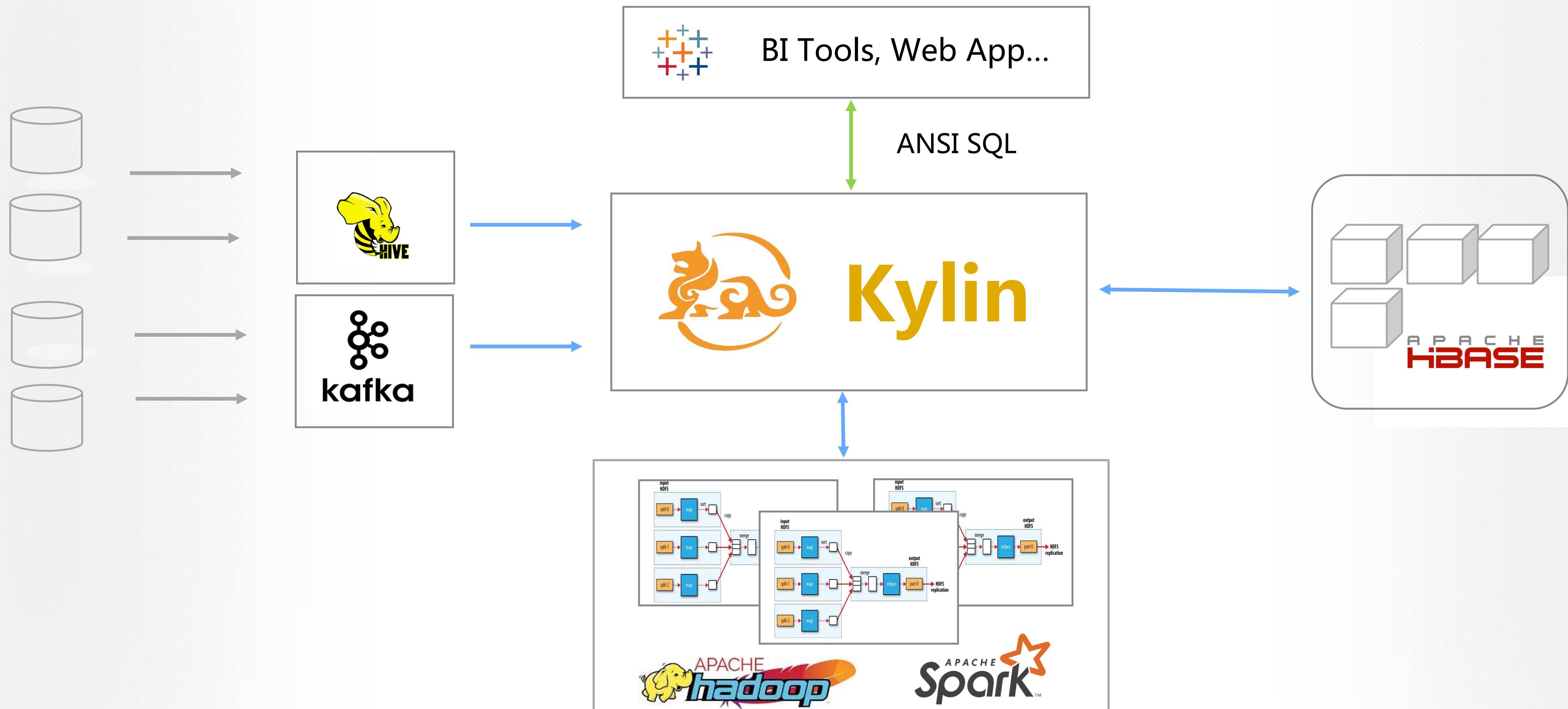
→ 预计算



Apache Kylin 系统架构

OSC 源创会
Opensource Innovation Meetup

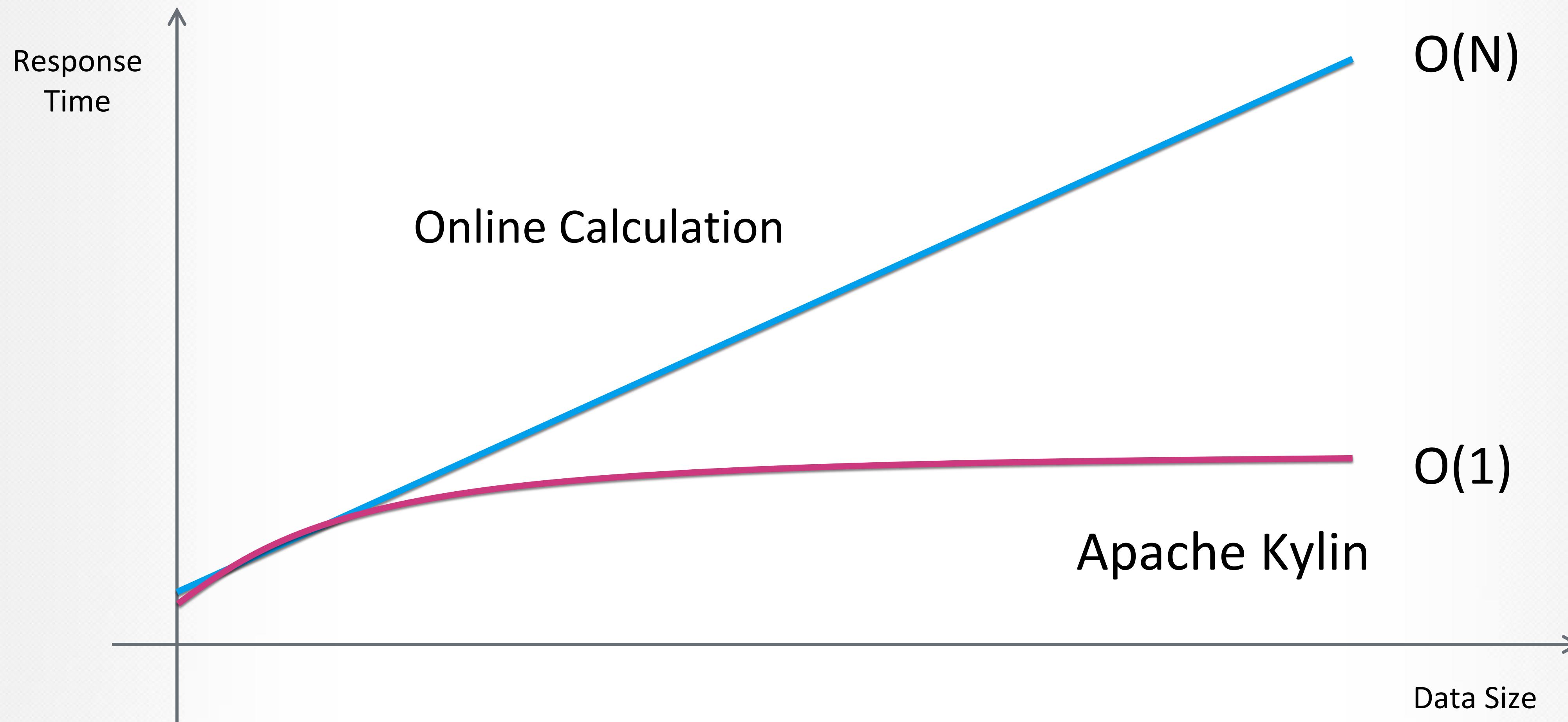
IT大咖说
知识分享平台



加速大数据OLAP分析

osc 源创会
Opensource Innovation Meetup

IT大咖说
知识分享平台



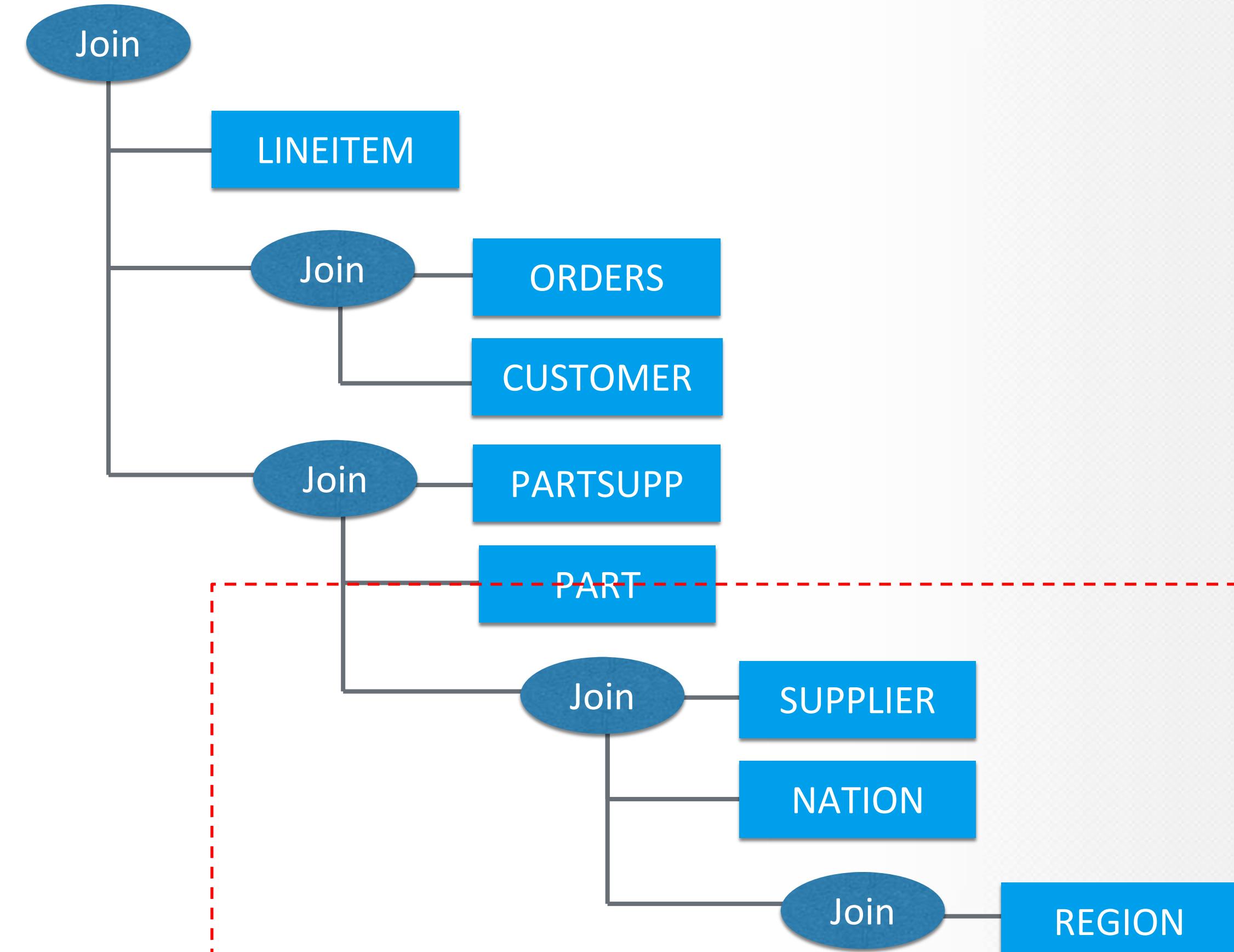
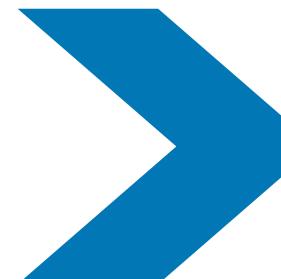
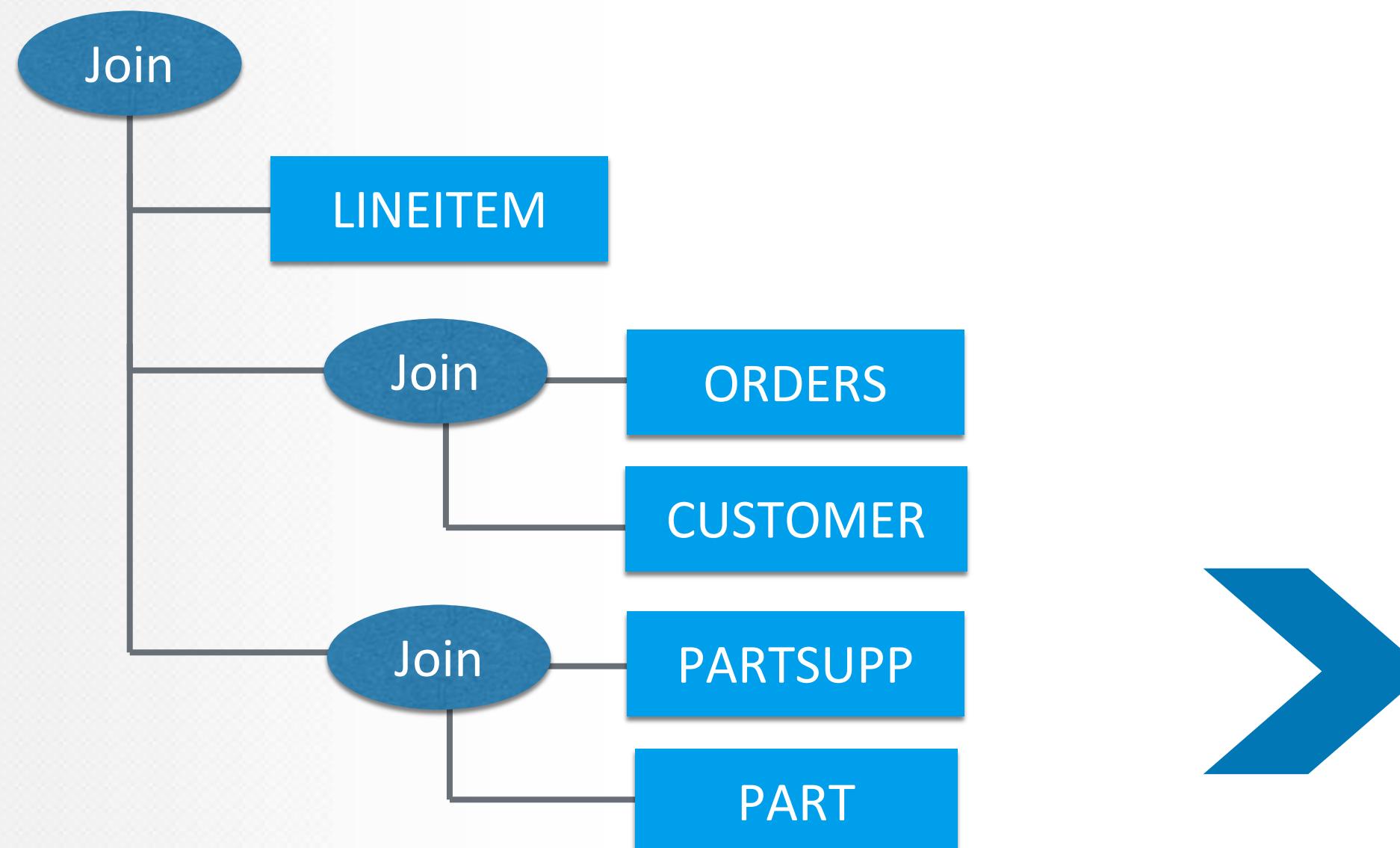
支持雪花模型

TPC-H Benchmarking

Kylin v2.0 支持雪花模型

OSC 源创会
Opensource Innovation Meetup

IT大咖说
知识分享平台



解决了Kylin 1.0很多功能限制：

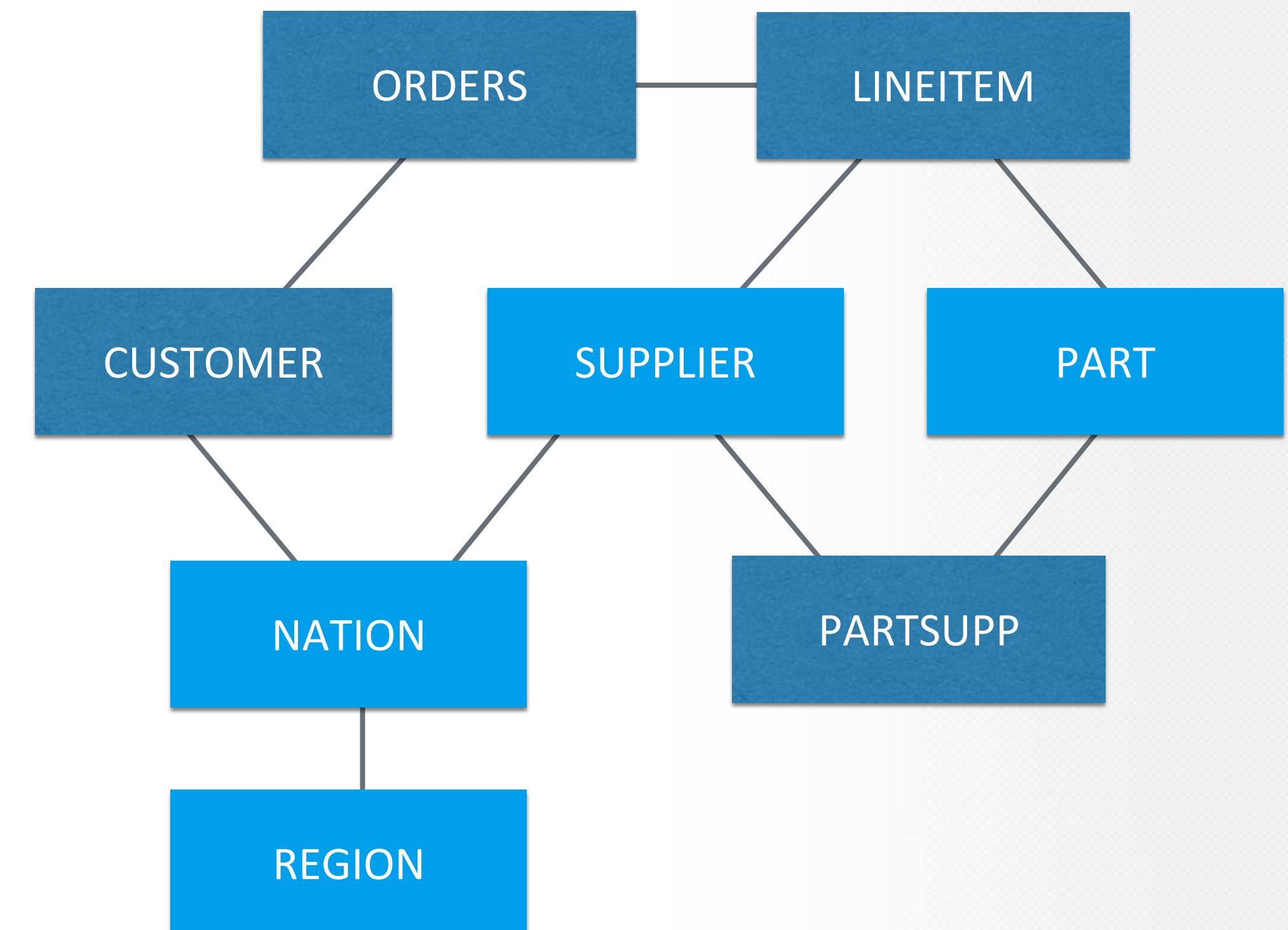
- 从星形模型到雪花模型
- 单表重复Join
-

TPC-H is a benchmark for decision support system.

- Popular among commercial RDBMS & DW solutions
- Queries and data have broad industry-wide relevance
- Examine large volumes of data
- Execute queries with a high degree of complexity
- Give answers to critical business questions

Kylin 2.0 runs all the 22 TPC-H queries. ([KYLIN-2467](#))

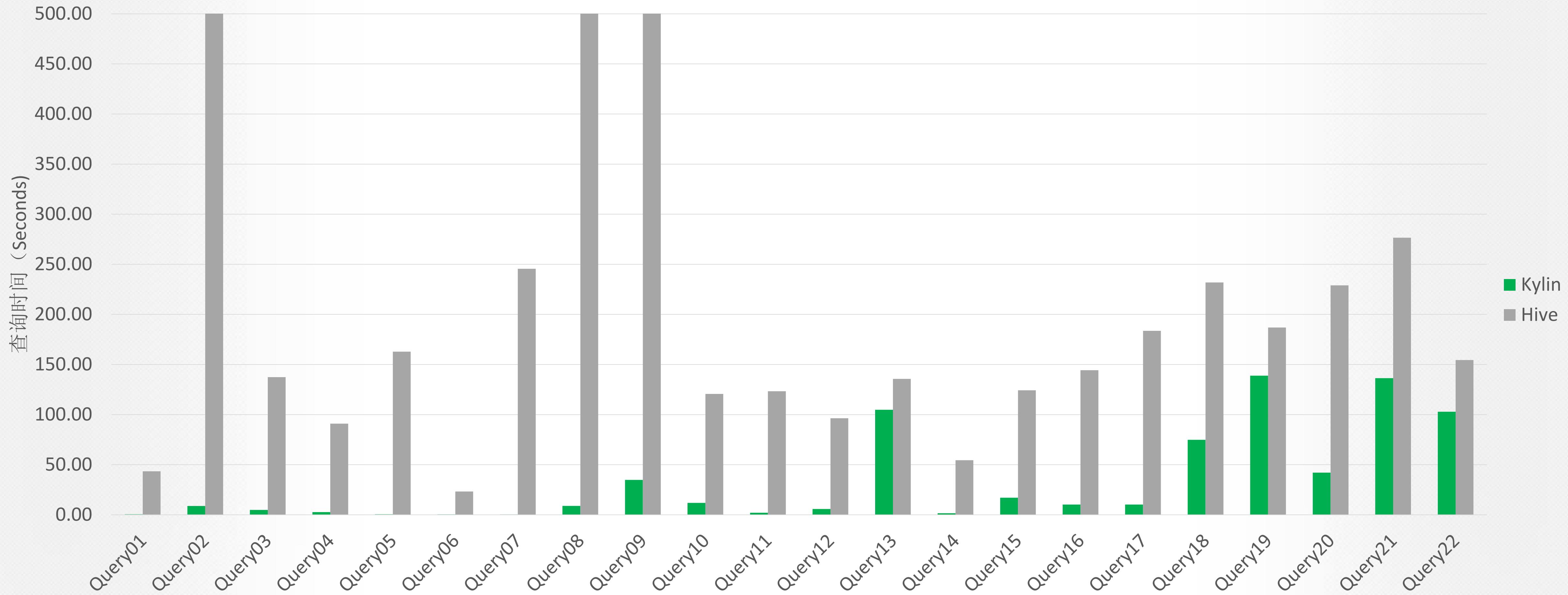
- Pre-calculation can answer very complex queries
- Goal is functionality at this stage
- Try it: <https://github.com/Kyligence/kylin-tpch>



Kylin vs Hive

OSC 源创会
Opensource Innovation Meetup

IT大咖说
知识分享平台

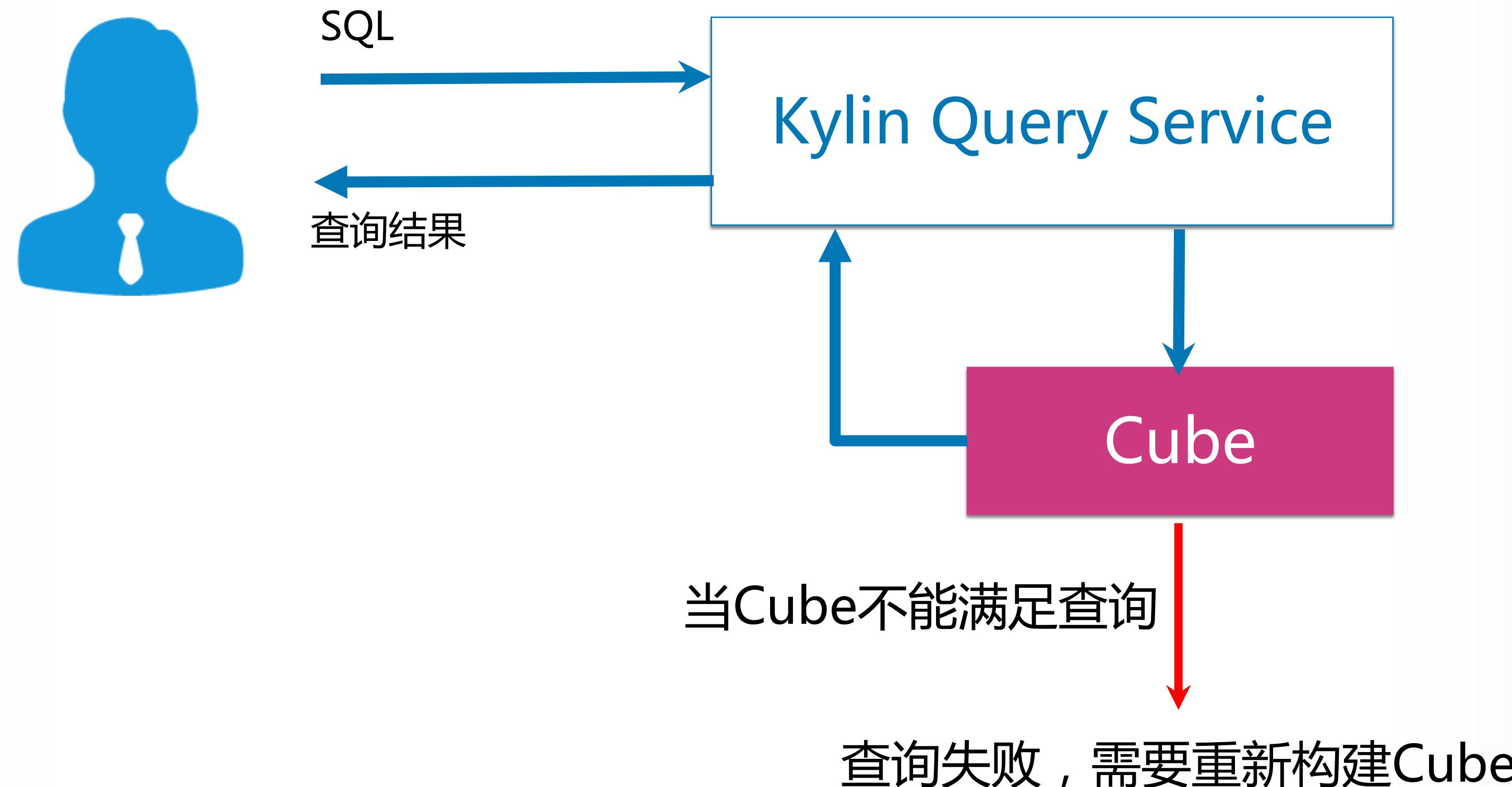


SQL Pushdown

v1.0 - 当Cube不能满足查询

OSC 源创会
Opensource Innovation Meetup

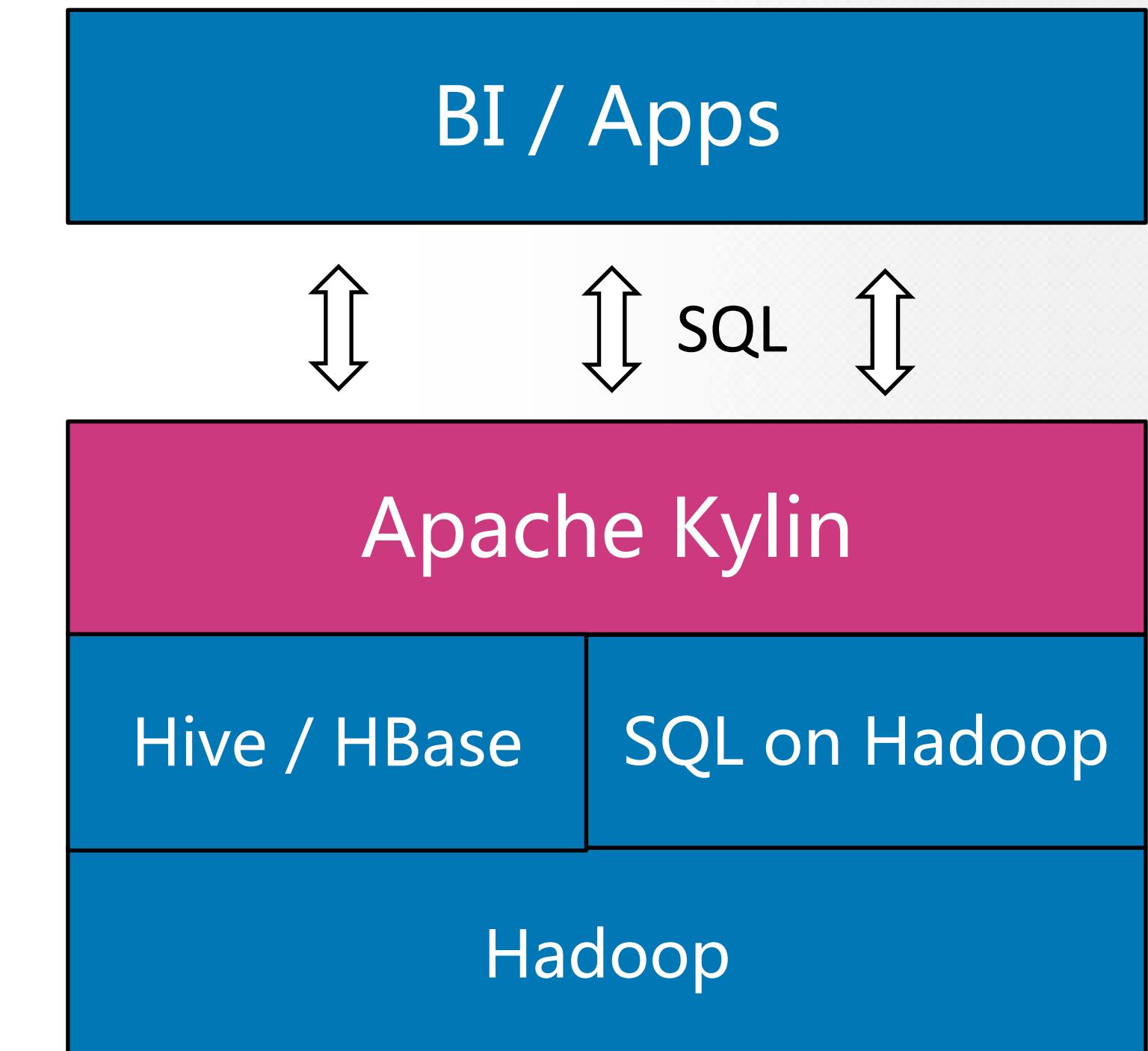
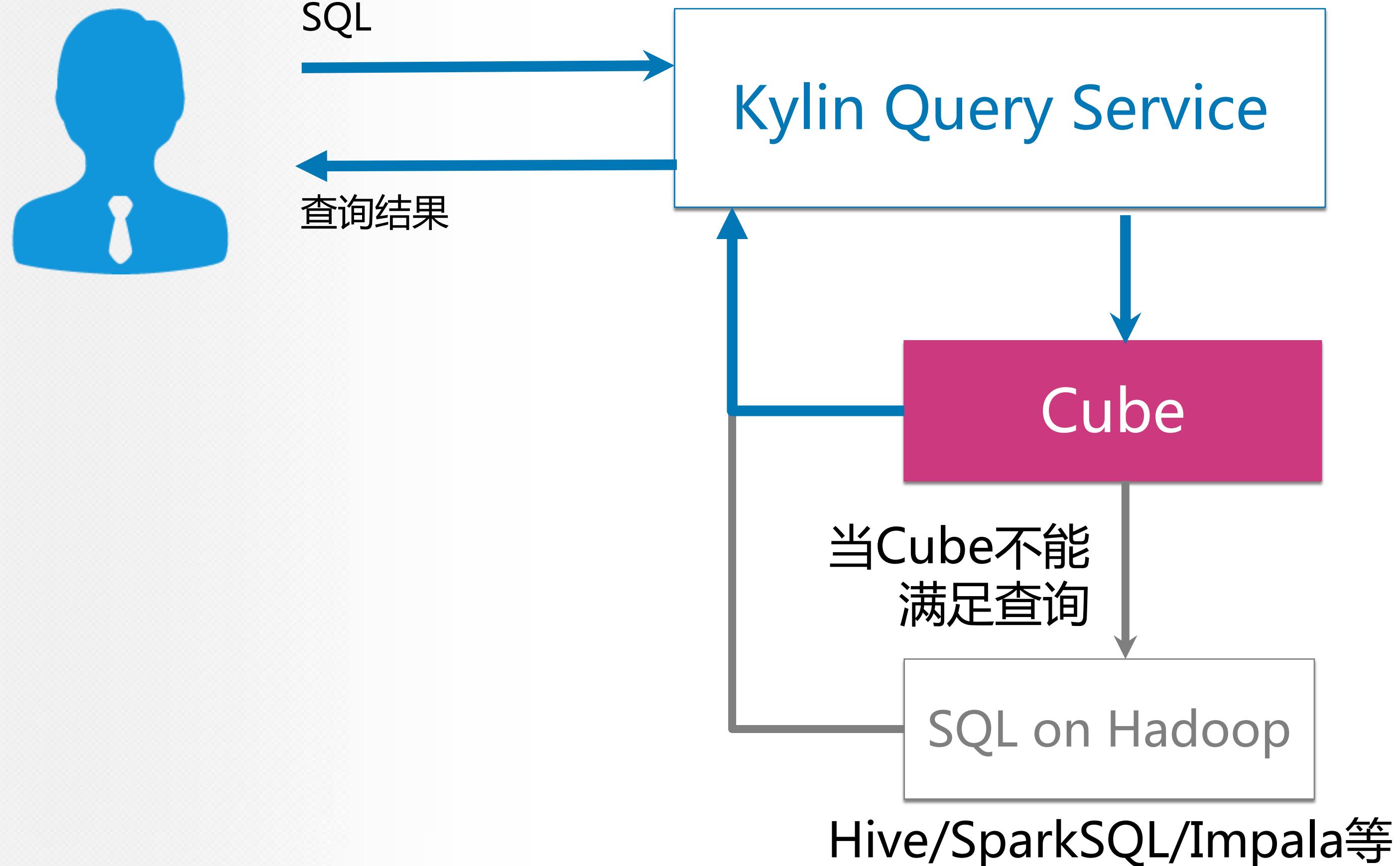
IT大咖说
知识分享平台



v2.1 - 当Cube不能满足查询

OSC 源创会
Opensource Innovation Meetup

IT大咖说
知识分享平台

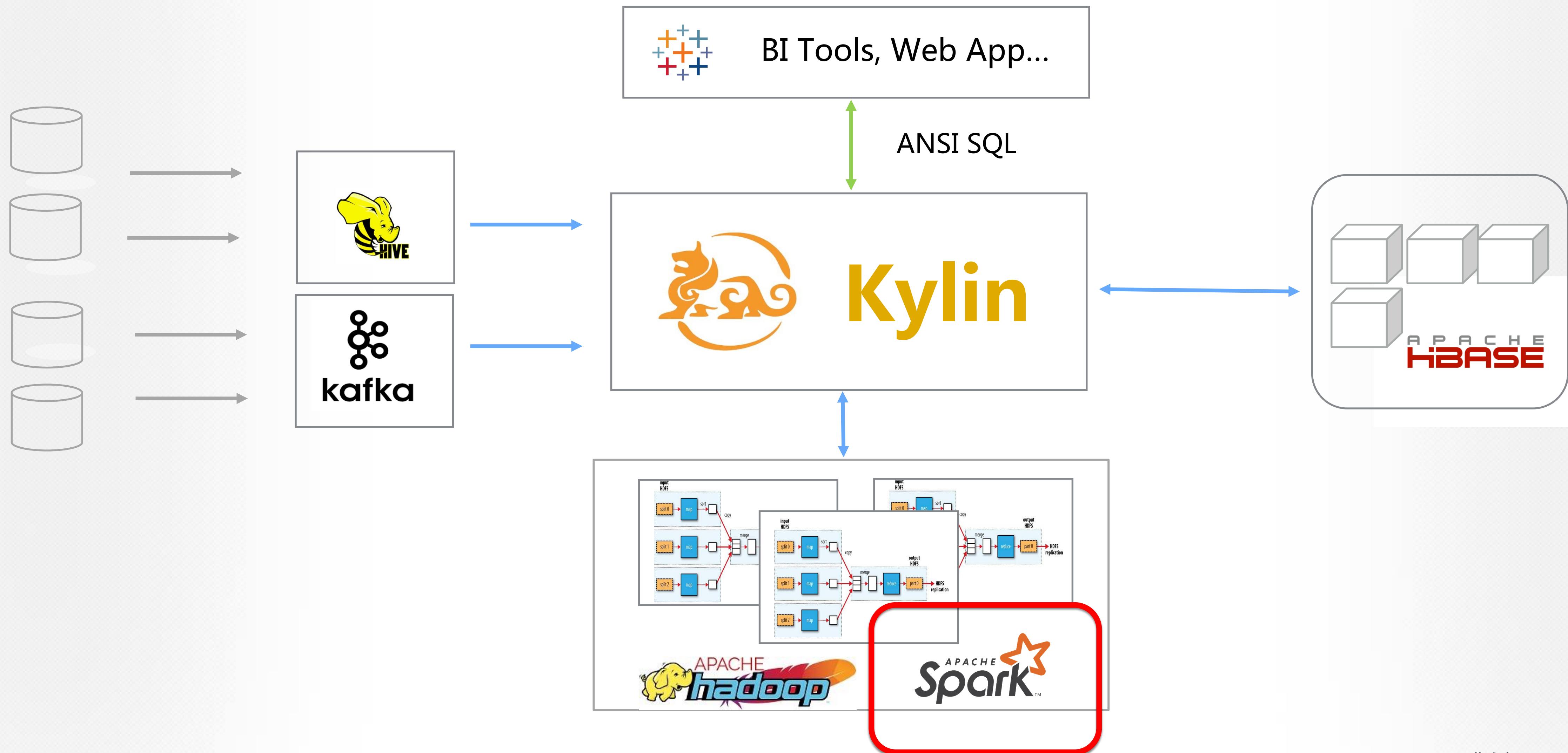


Spark Cubing

Cubing with Spark

osc 源创会
Opensource Innovation Meetup

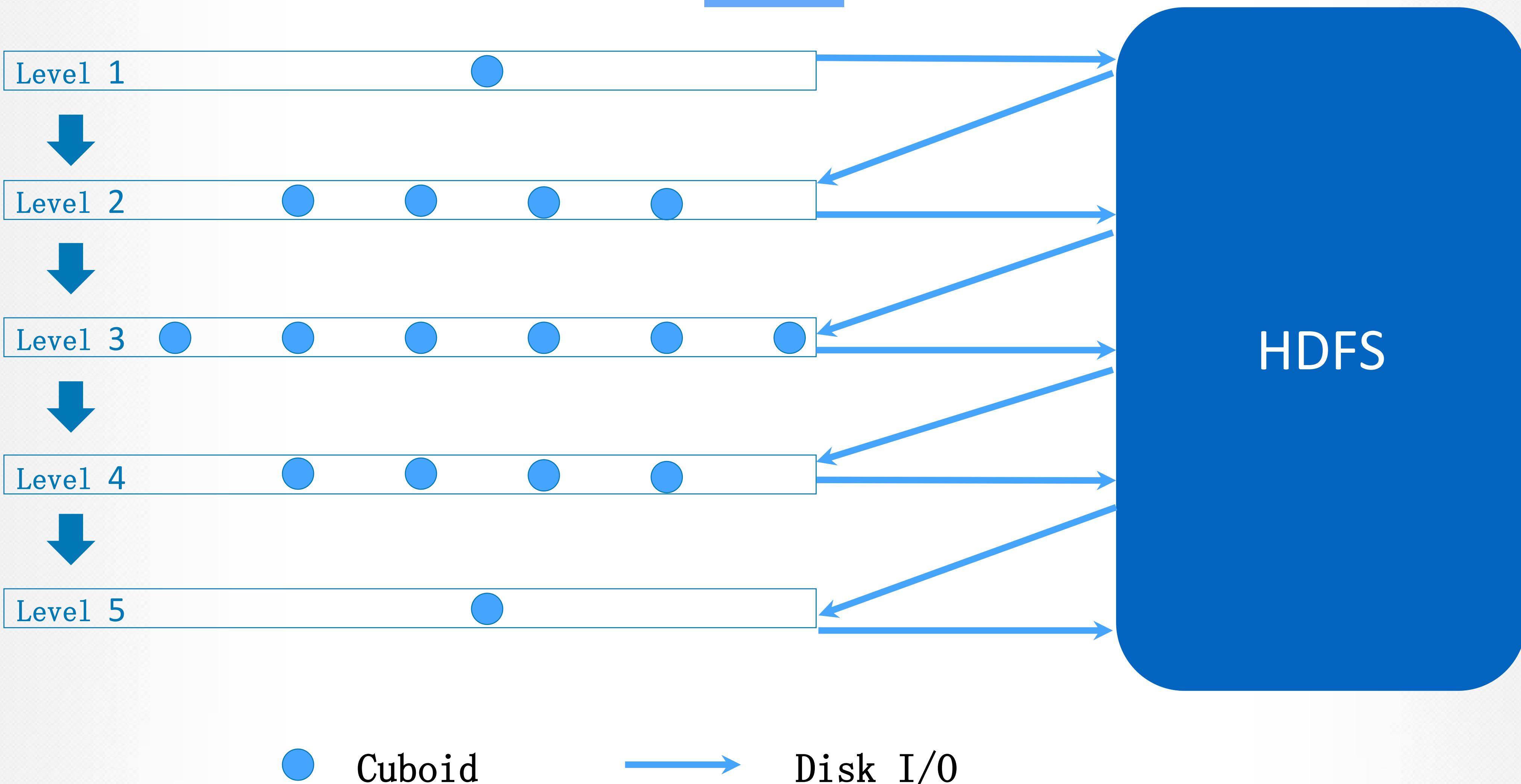
IT大咖说
知识分享平台



MR Layered Cubing

OSC 源创会
Opensource Innovation Meetup

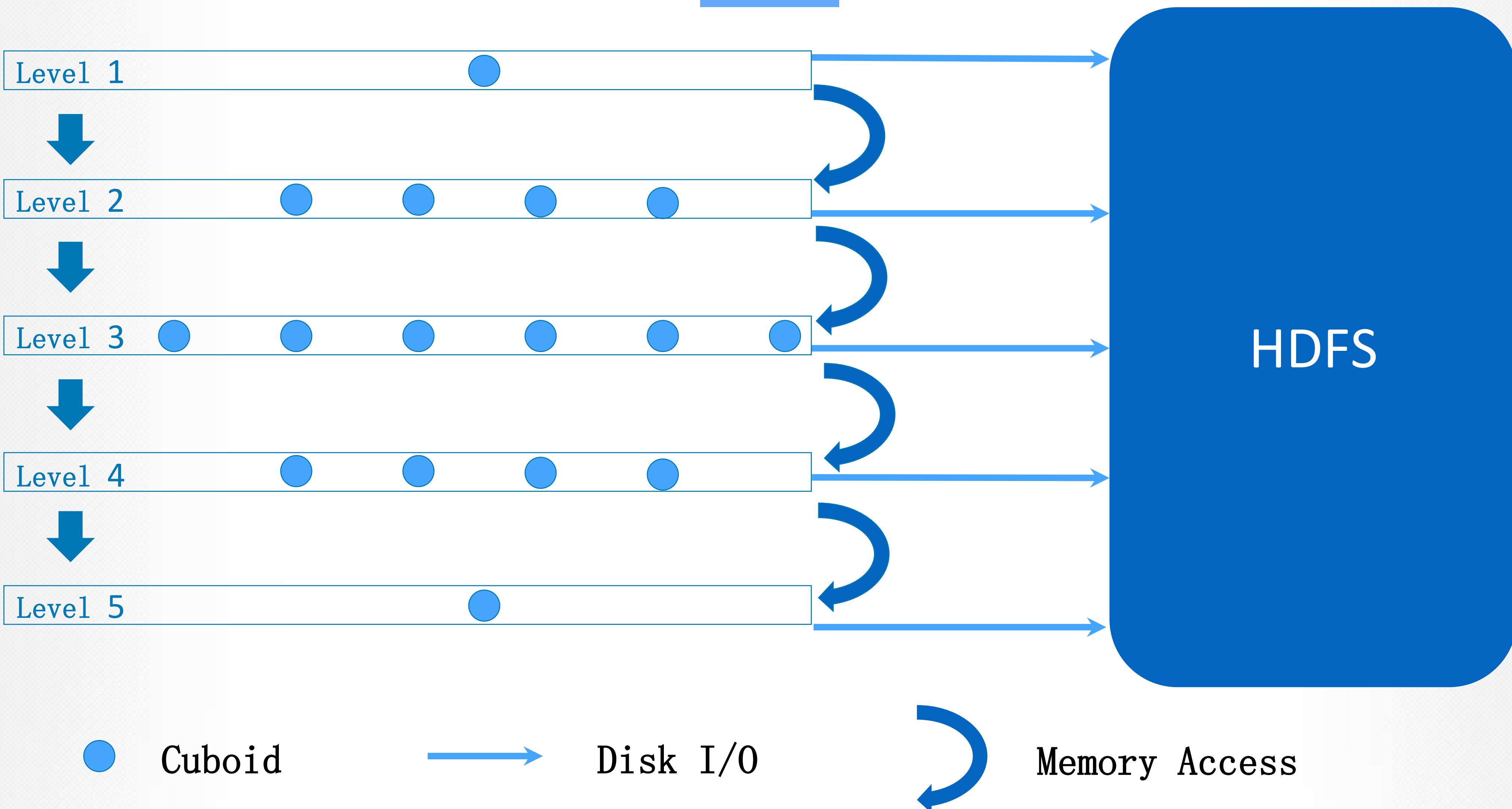
IT大咖说
知识分享平台



Spark Cubing

OSC 源创会
Opensource Innovation Meetup

IT大咖说
知识分享平台



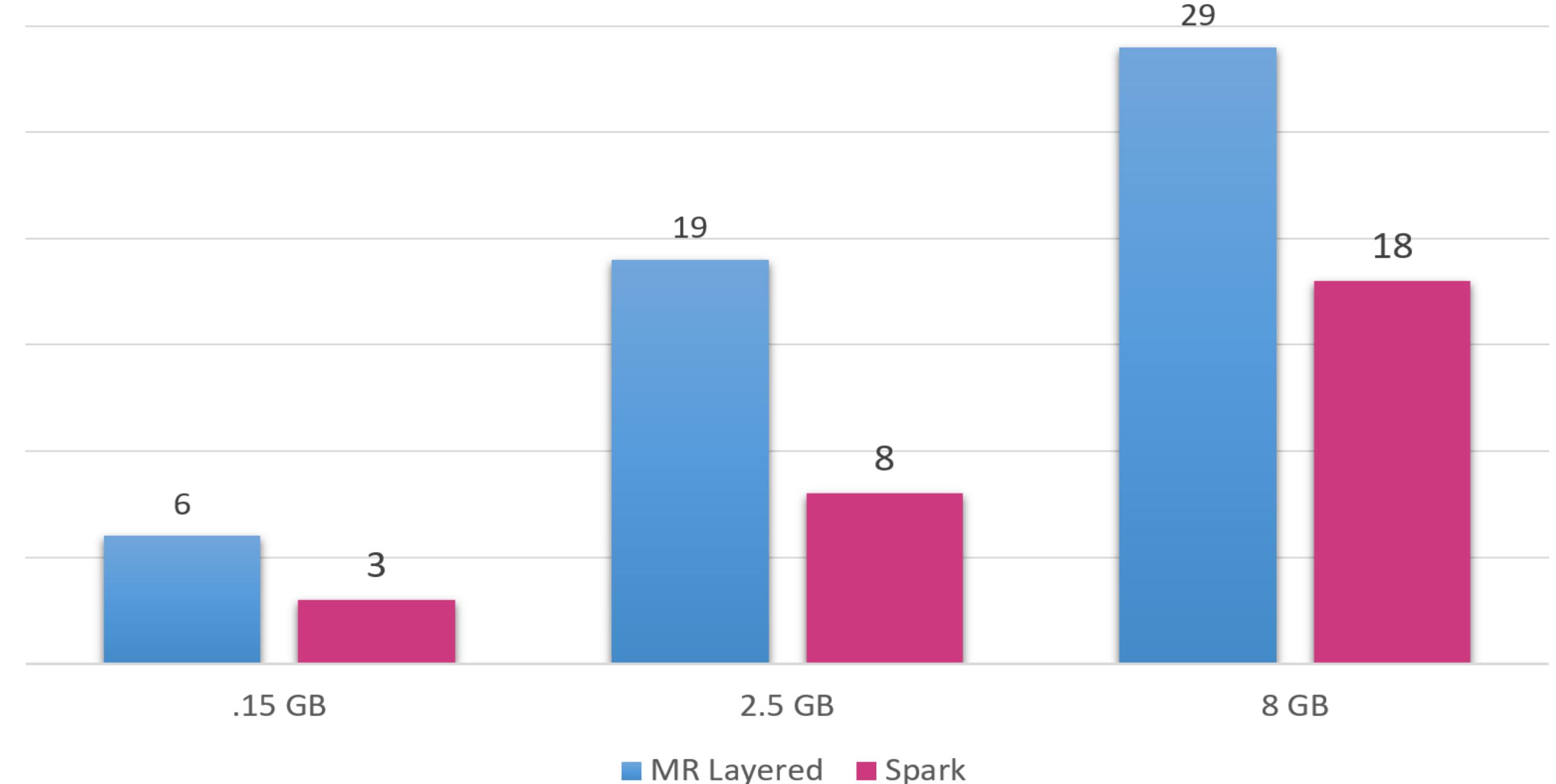
Spark Cubing vs. MR Layered

Scubing 源创会
Opensource Innovation Meetup

IT大咖说
知识分享平台

构建时间减半，但是可以观察到优势随着数据量的增加而减少

- 4 节点的集群
- Spark 1.6.3 on YARN
- 24 vcores, 30 GB memory
- 3 data sets of increasing size:
.15 GB / 2.5 GB / 8 GB



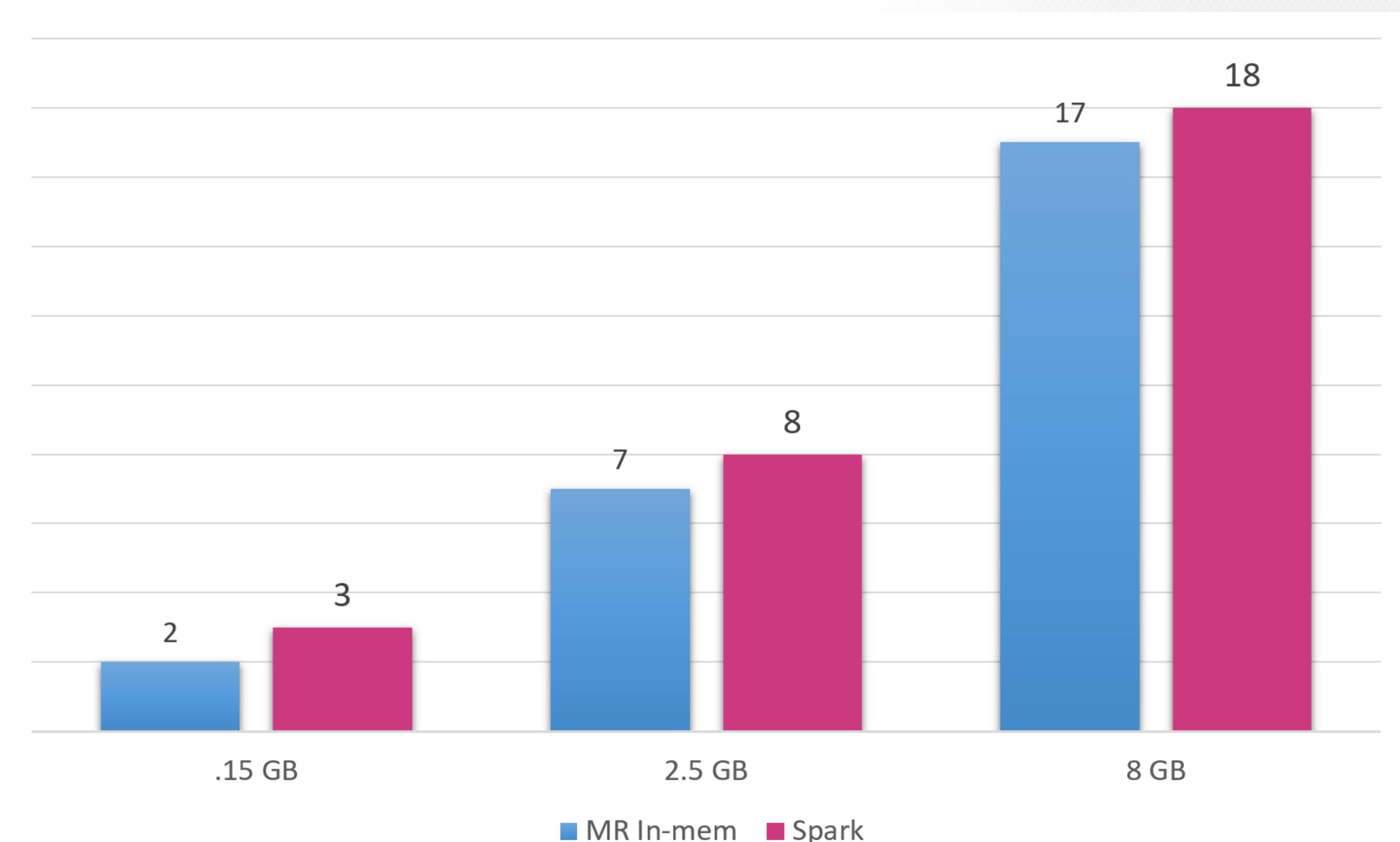
Spark Cubing vs. MR In-mem

Scubing 源创会
Opensource Innovation Meetup

IT大咖说
知识分享平台

几乎一样快，但是更适合通用的数据集

- In-mem cubing 期望良好分区的数据，在随机分布的数据上表现次优
- Spark cubing 更适合随机分布的数据



Can be cooler ?

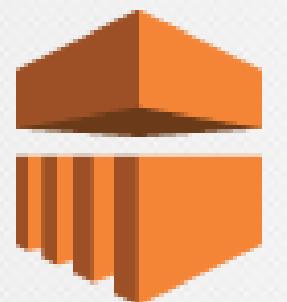
云端一键部署

OSC 源创会
Opensource Innovation Meetup

IT大咖说
知识分享平台



Azure HDInsight



AWS EMR

.....

Kyligence Cloud !

KAP on HDInsight

Search by app name or ID

Available applications

- Kyligence Analytics by KYLIGENCE
- StreamSets Data Collector by STREAMSETS

Offer details

Kyligence Analytics Platform 2.3 by Kyligence

Terms of use | privacy policy

Terms of use

By clicking "Purchase", I (a) agree to the legal terms and privacy statement(s) associated with each marketplace offering above, (b) authorize Microsoft to charge or bill my current payment method for the fees associated with my use of the offering(s), including applicable taxes, with the same billing frequency as my Azure subscription, until I discontinue use of the offering(s), (c) agree that Microsoft may share my contact information and transaction details (including usage volume associated with the offering) with the seller(s) of the offering(s), and (d) give Microsoft permission to share my contact information so that the provider of the offer can contact me regarding this product and related products. Microsoft does not provide rights for third-party products or services. See the Azure Marketplace Terms for additional terms.

Unavailable applications

The following applications are not compatible with your cluster configuration.

- CDAP 4.1 for HDInsight
- CDAP 4.2 for HDInsight
- Datameer
- DSS on HDInsight
- H2O Artificial Intelligence for HDInsight
- SnappyGzip Hadoop
- Spark Job Server for KNIME Spark

Kyligence Cloud Beta

Dashboard / cluster

+Cluster

Name	Status	Create Time	Action
testkap24	CREATING	2017-08-29 13:54:13	★Start
ssb	RUNNING	2017-09-21 18:35:10	★Start

Basic Info

Cluster Name: ssb Status: RUNNING

Cluster Type: EMR Create Time: 2017-09-21 18:35:10

Topology: SINGLE

Cluster START STOP RESIZE

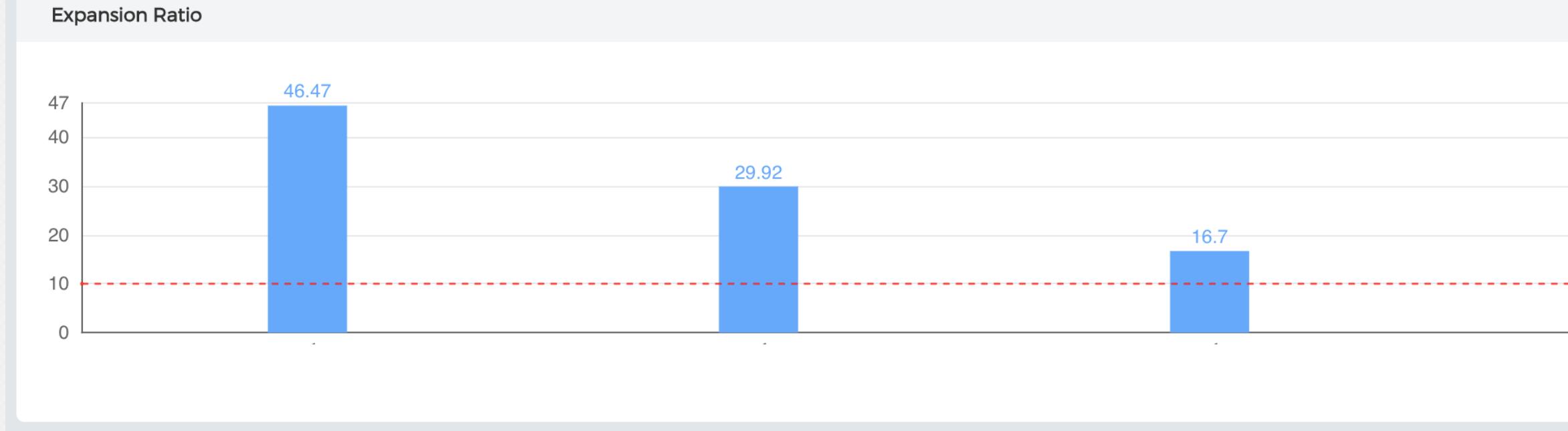
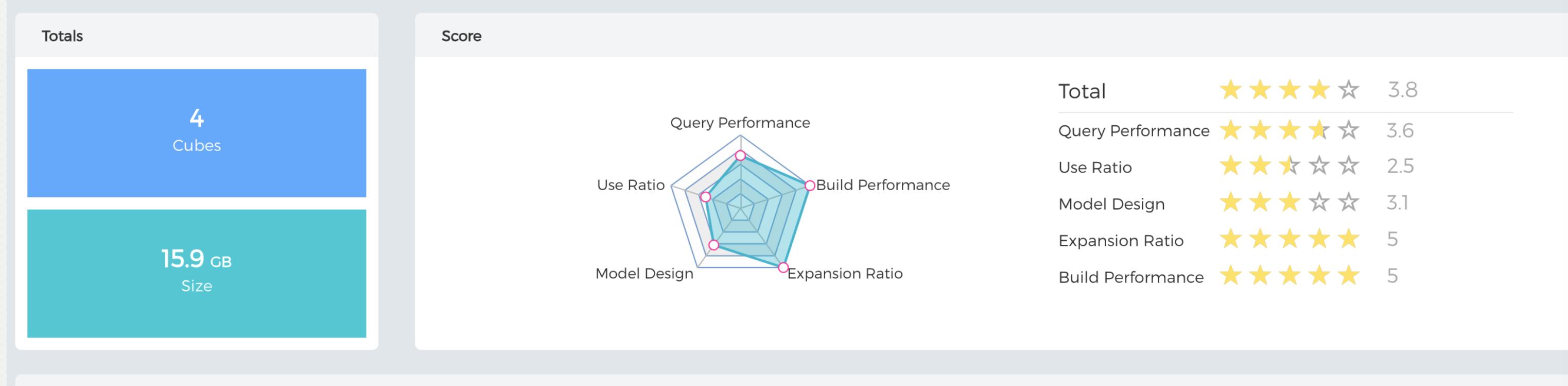
Cluster Type: Status: RUNNING

Work Node Count: 1 Create Time: 2017-09-21 18:35:10

KAP

Total 2 10 /page < 1 > Go to 1

自助式 Cube 调优



Kyligence Robot !

<https://kybot.io>

- 统计评分
- 优化建议

Rowkey

Suggestion Origin

Basic Information

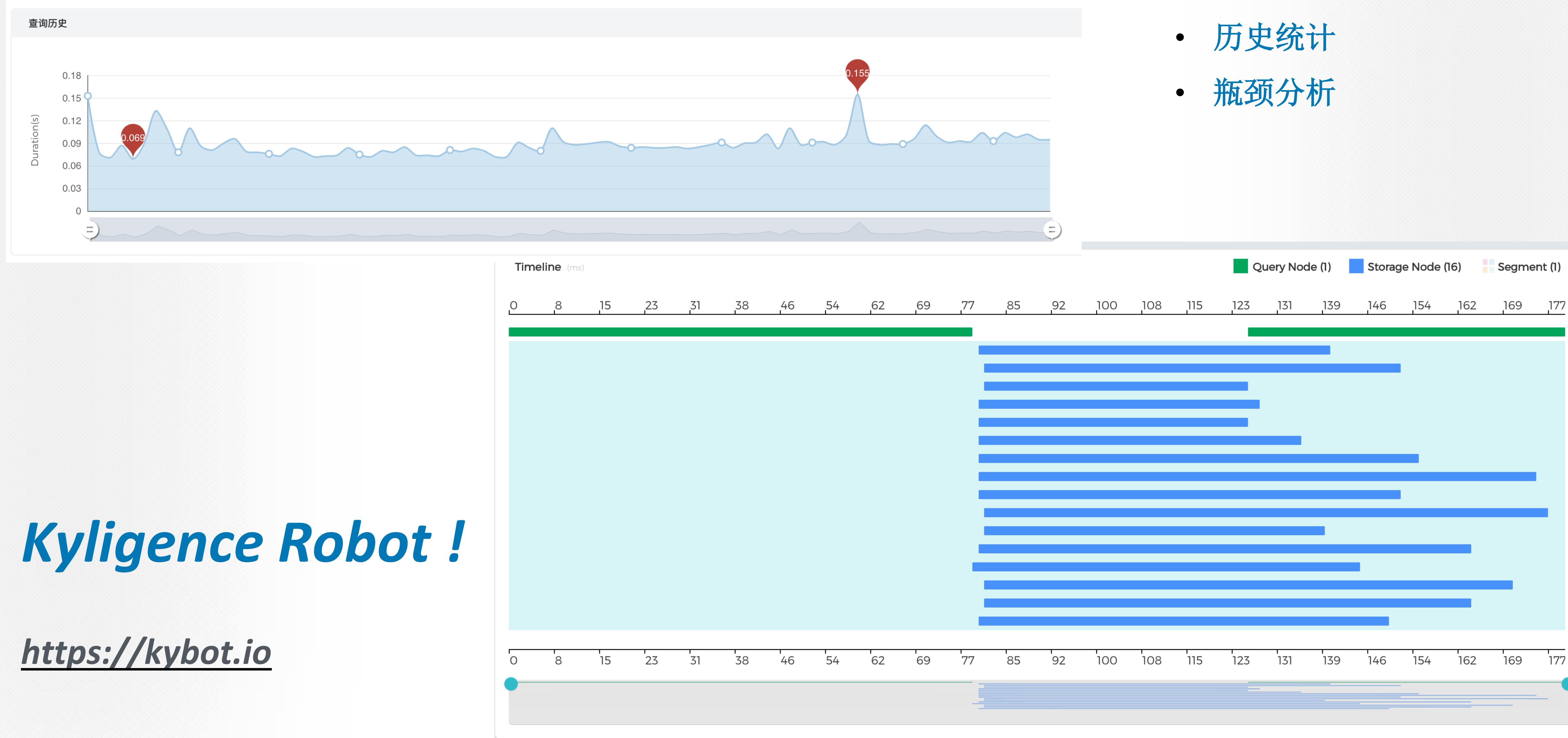
Column Type Encoding

Rowkey	Cardinality	Appear	Absent	difference with parent cuboid
CATEGORY	1007	varchar(256)	dict	
COUNTRY_CODE	9	integer	integer:4	
PROVINCE_CODE	8	integer	integer:4	
CITY	6	varchar(256)	fixed_length:100 dict	
POSTCODE	6	varchar(256)	fixed_length:120 dict	
AREA	5	varchar(256)	fixed_length:100 dict	
GENDER	2	varchar(256)	fixed_length:40 dict	
IS_VALID	2	varchar(256)	fixed_length:40 dict	
IS_ONLINE	2	varchar(256)	dict	
IS_ACTIVITY	2	varchar(256)	dict	

自助式查询优化

OSC 源创会
Opensource Innovation Meetup

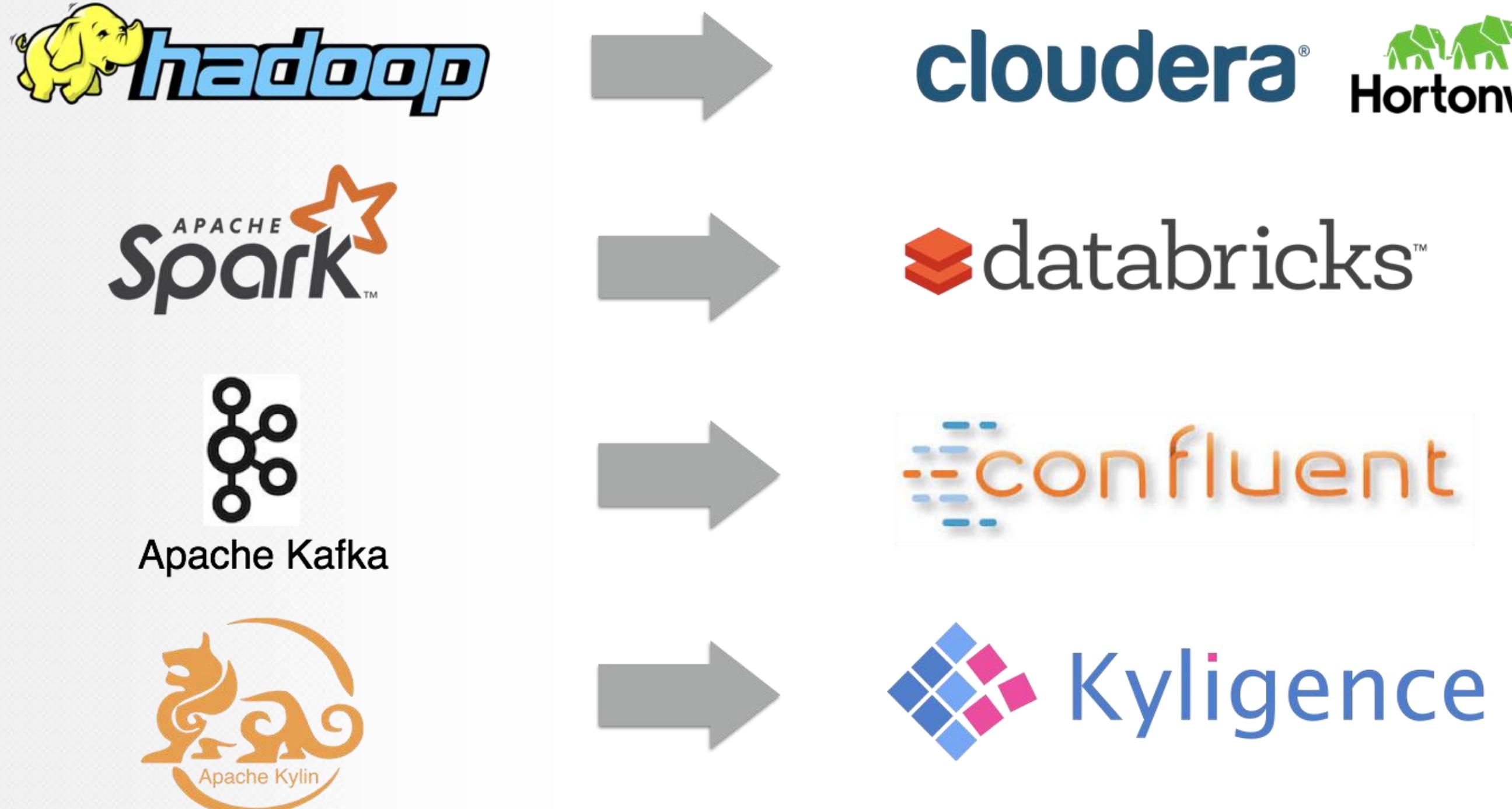
IT大咖说
知识分享平台



关于Kyligence

OSC 源创会
Opensource Innovation Meetup

IT大咖说
知识分享平台



构建领先的
全球开源社区

Apache Kylin v2.2 is released!

<http://kylin.apache.org>

We are hiring!

OSC 源创会 | IT 大咖说
Opensource Innovation Meetup

谢谢！

李栋

