OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识分享平台

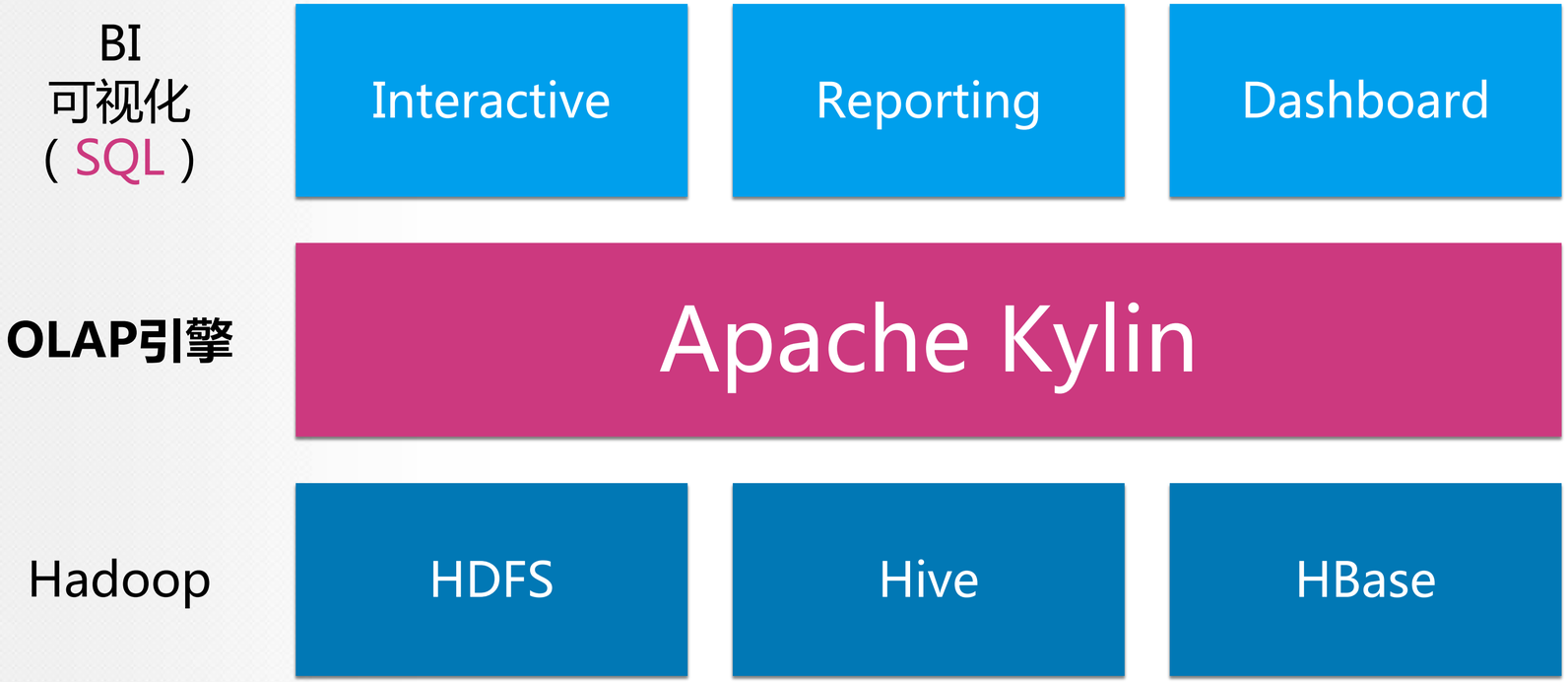
Unleash big data productivity

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？



- 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

Sep 2013 项目开始

Oct 2014 正式开源

Nov 2014 加入Apache孵化器项目

Sep 2015 首个来自中国的Apache顶级项目

Nov 2015 InfoWorld: Bossie Award 最佳开源大数据工具奖

Mar 2016 Kyligence公司创建

Aug 2016 商业版KAP发布

Sep 2016 InfoWorld: Bossie Award 最佳开源大数据工具奖

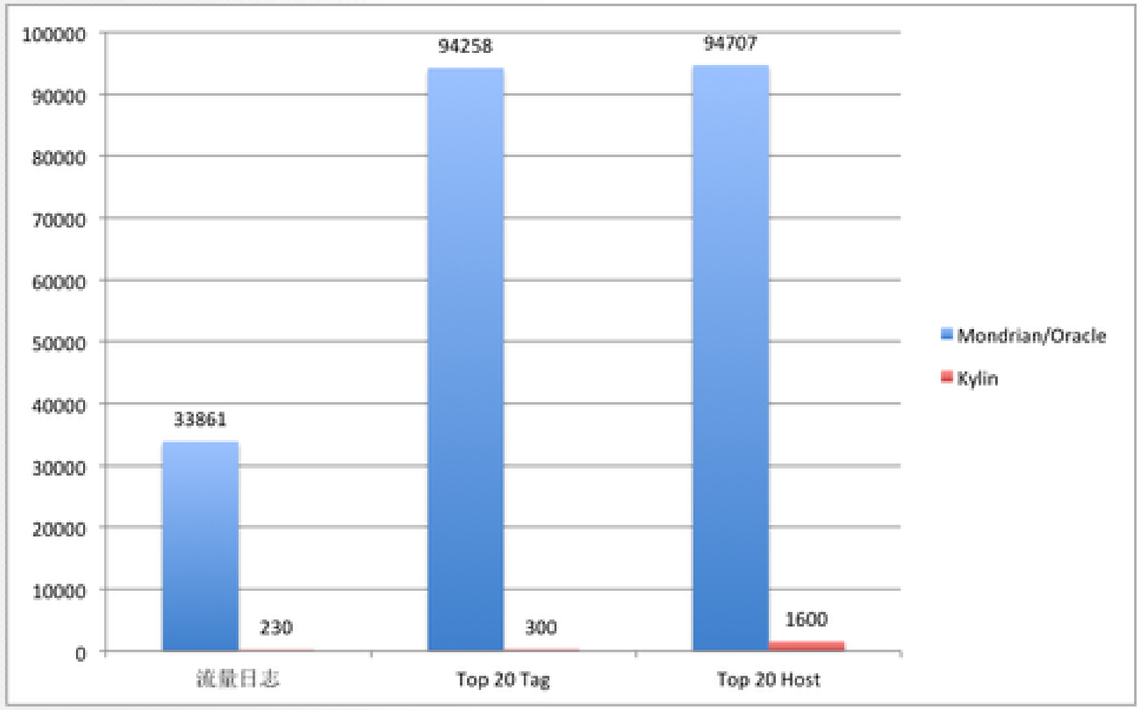
Nov 2017 Apache Kylin v2.2 发布

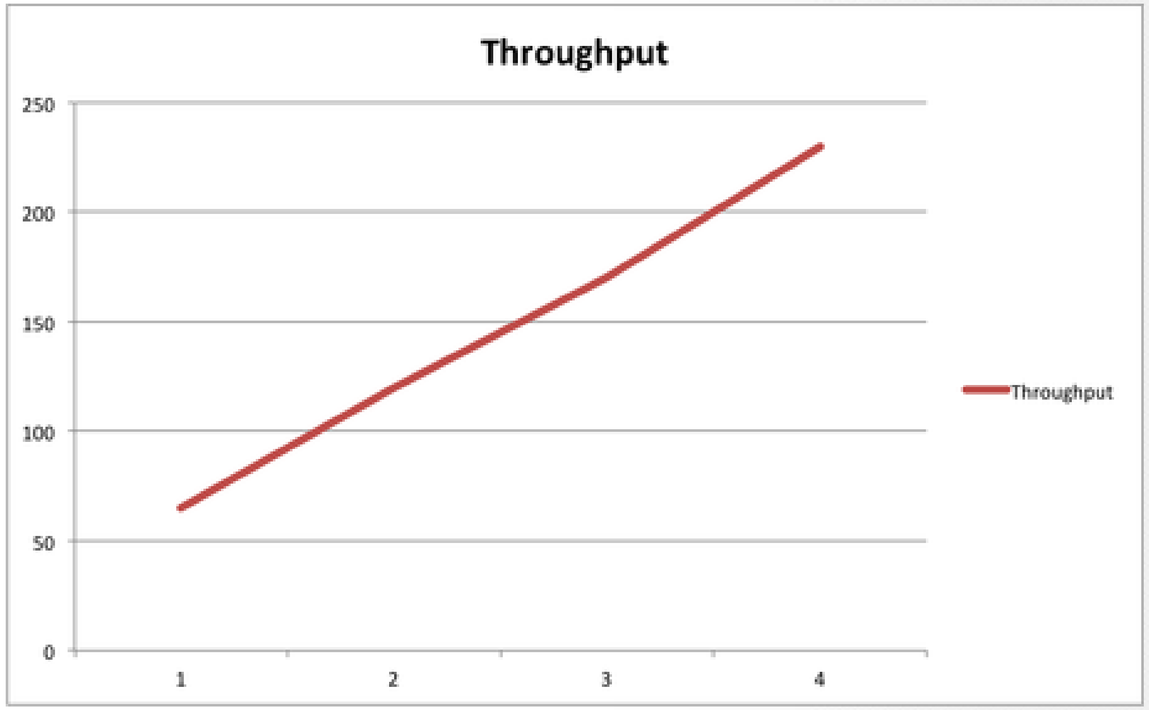
Apache Kylin全球案例

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

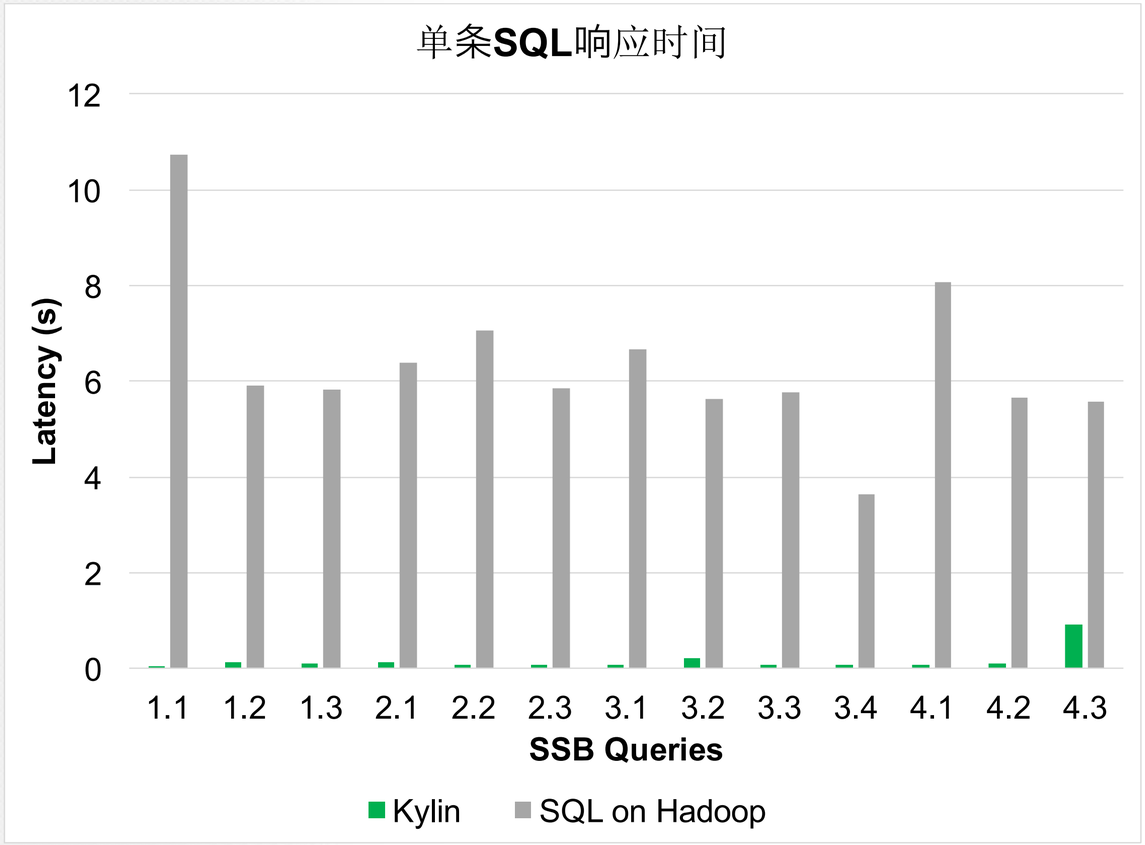
Apache Kylin性能测试

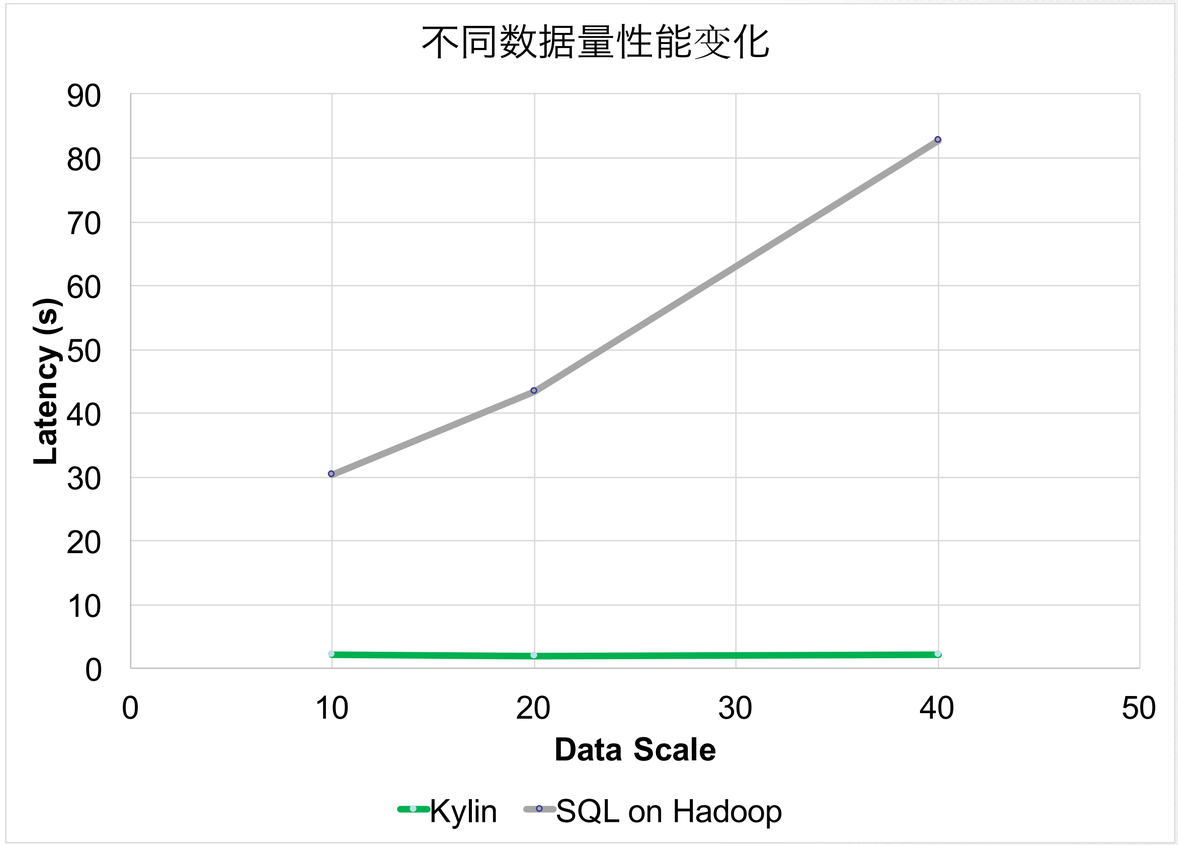




By 网易:
http://datalab.int-yt.com/archives/1708

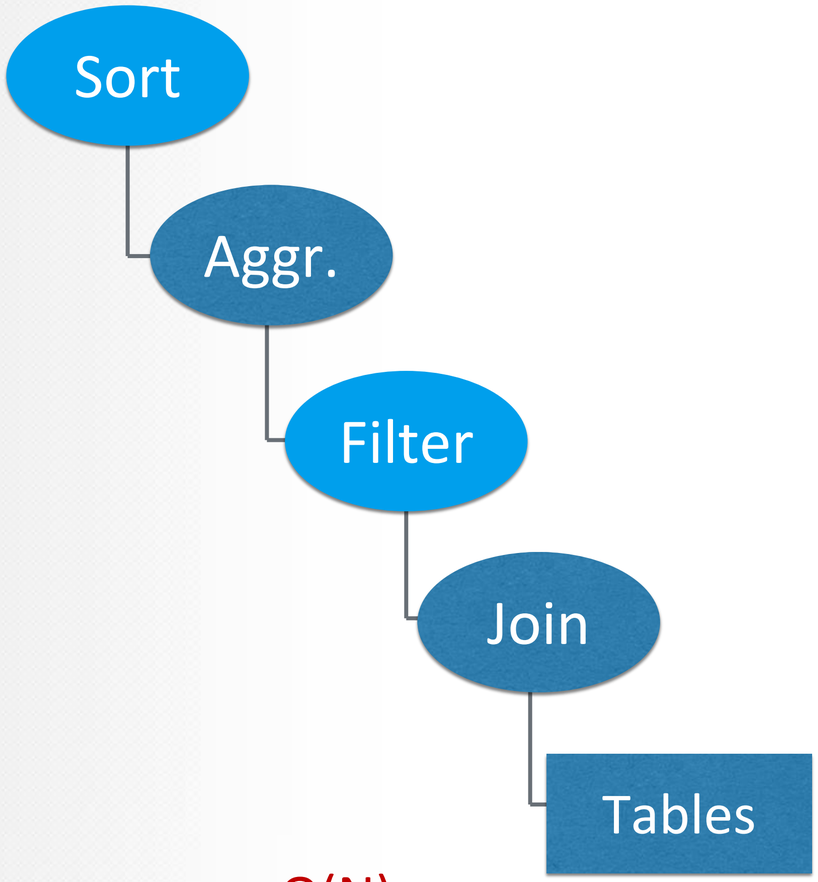
Apache Kylin vs. SQL-on-Hadoop





Star schema benchmark:
http://www.cs.umb.edu/~poneil/StarSchemaB.PDF

Apache Kylin 为什么快？

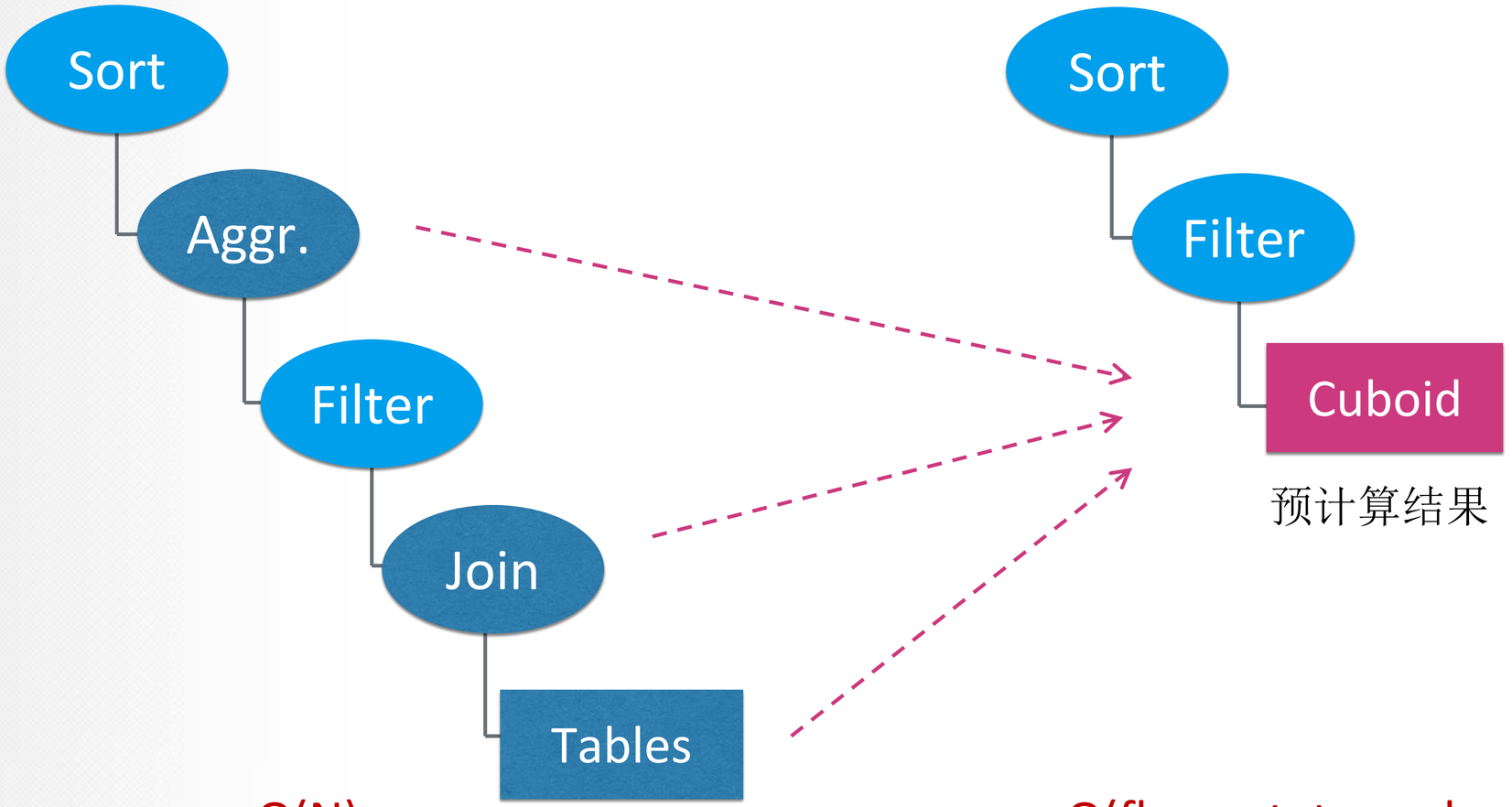


样例：

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

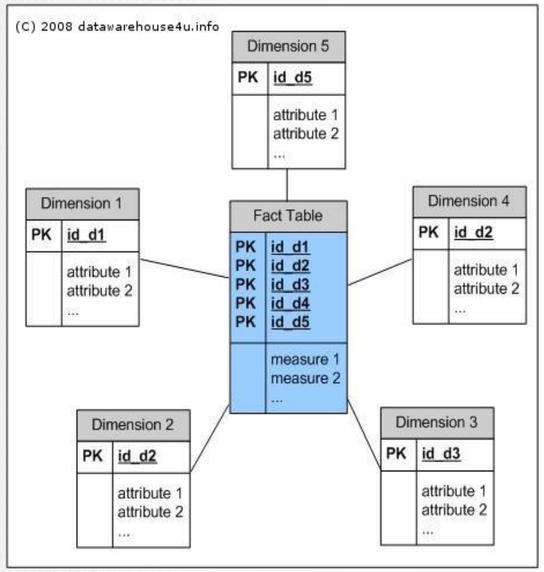
```
select
   l_returnflag,
   o_orderstatus,
   sum(l_quantity) as sum_qty,
   sum(l_extendedprice) as sum_base_price
   …
from
   v_lineitem
   inner join v_orders on l_orderkey = o_orderkey
where
   l_shipdate <= '1998-09-16'
group by
   l_returnflag,
   o_orderstatus
order by
   l_returnflag,
   o_orderstatus;
```

Apache Kylin 为什么快？

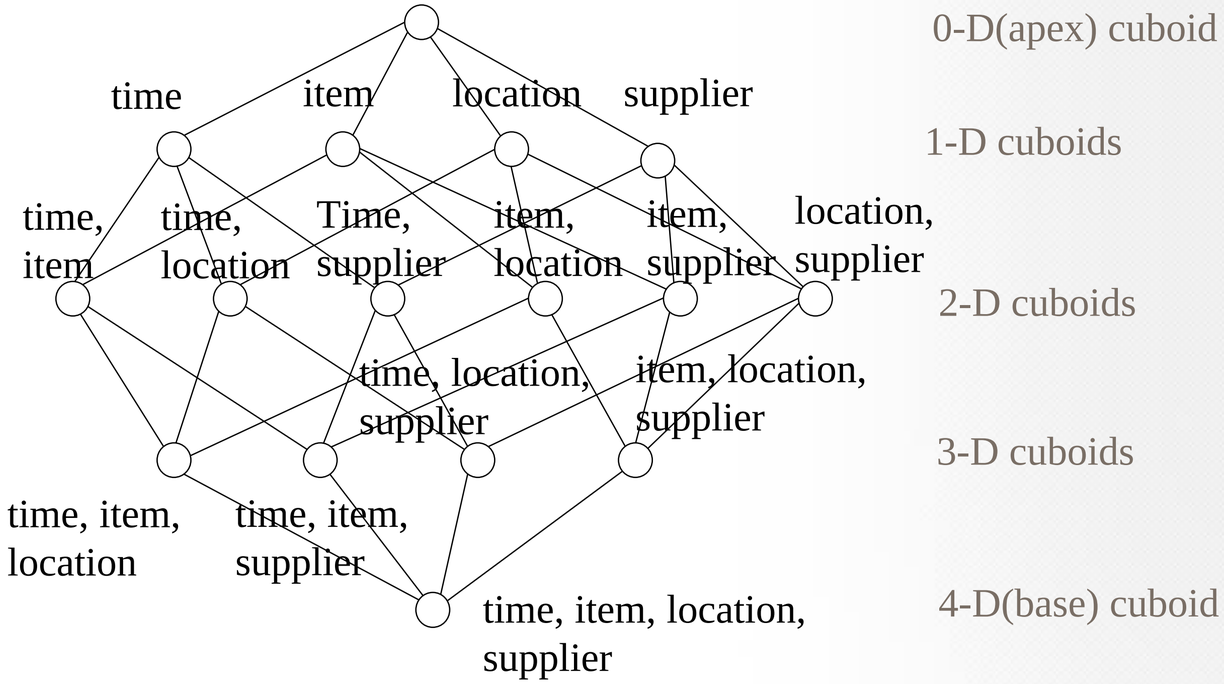


Apache Kylin 关键在于预计算

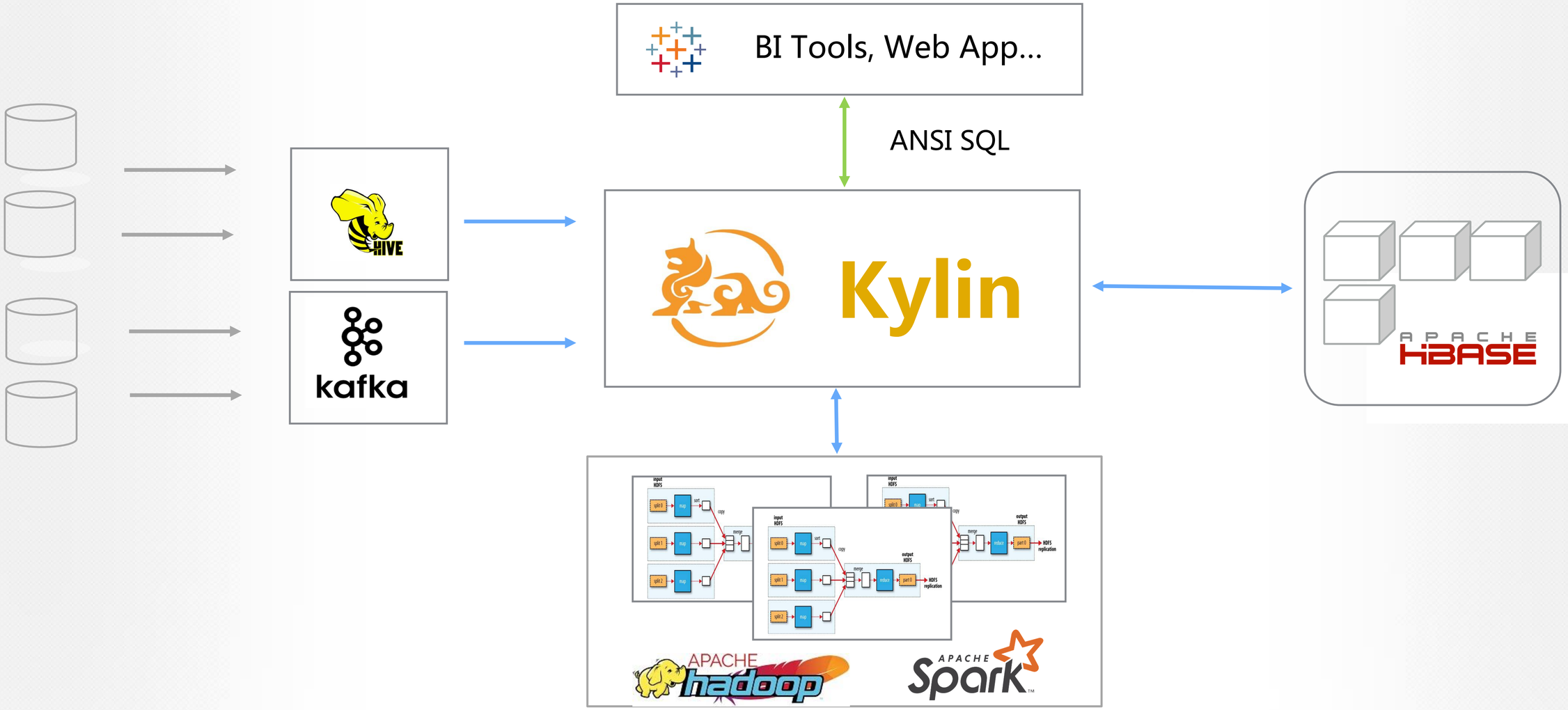
- OLAP Cube 理论基础
- Model 和 Cube 定义预计算范围
- Build Engine 执行预计算任务
- Query Engine 在预计算结果上完成查询



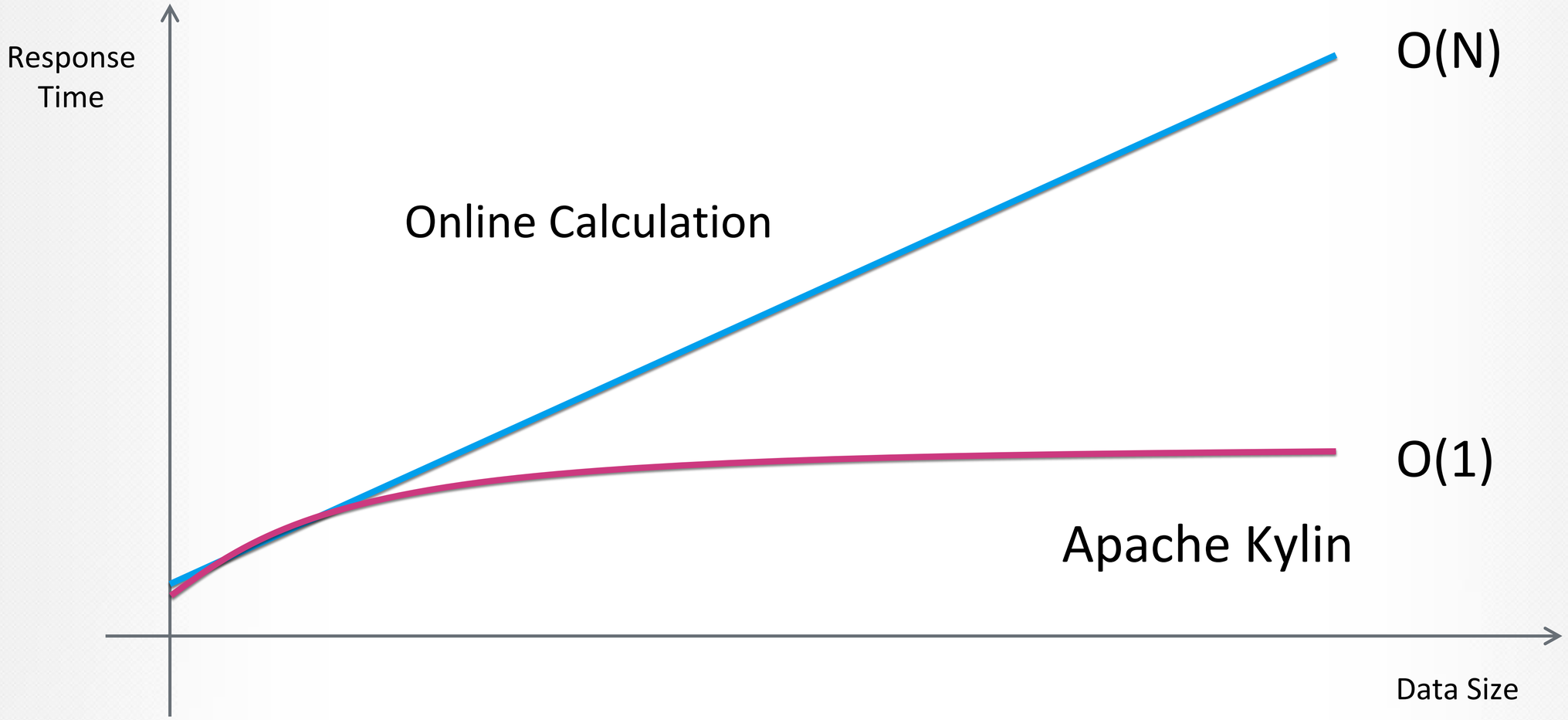
预计算



Apache Kylin 系统架构



加速大数据OLAP分析

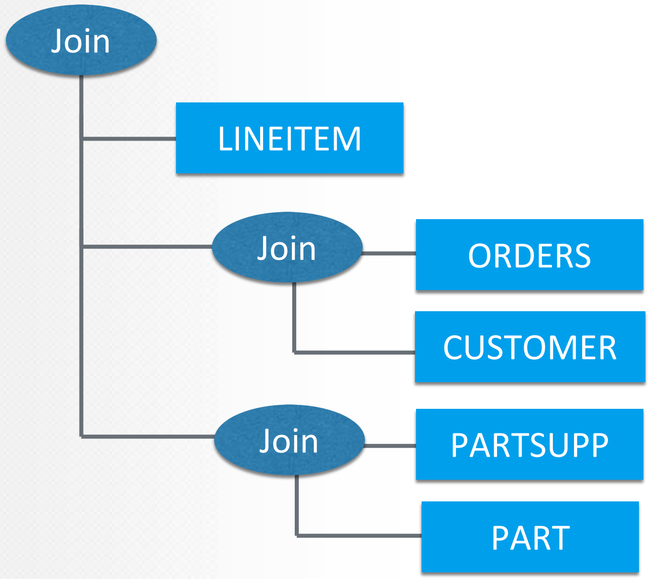


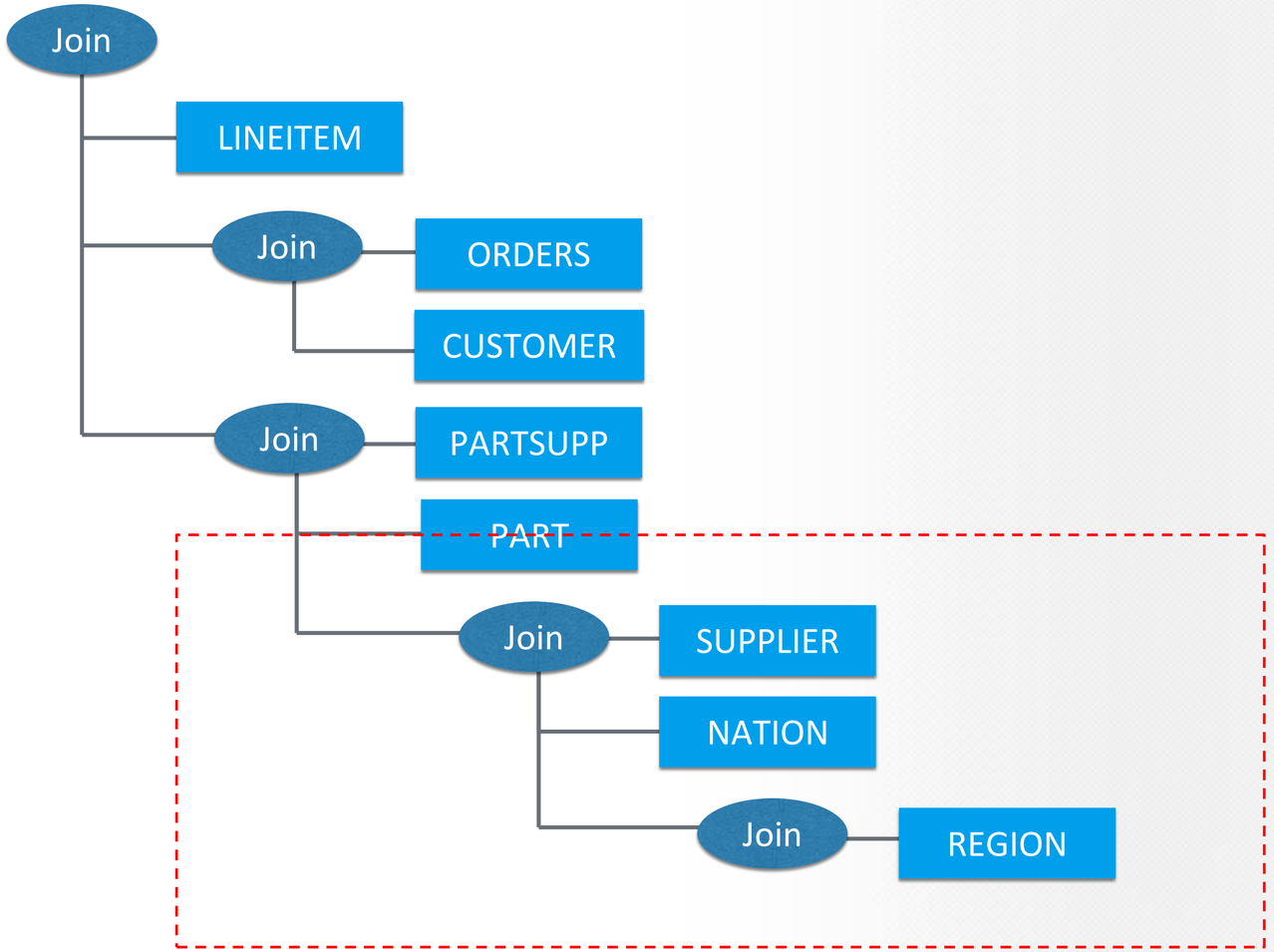
OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识分享平台

支持雪花模型

TPC-H Benchmarking

Kylin v2.0 支持雪花模型





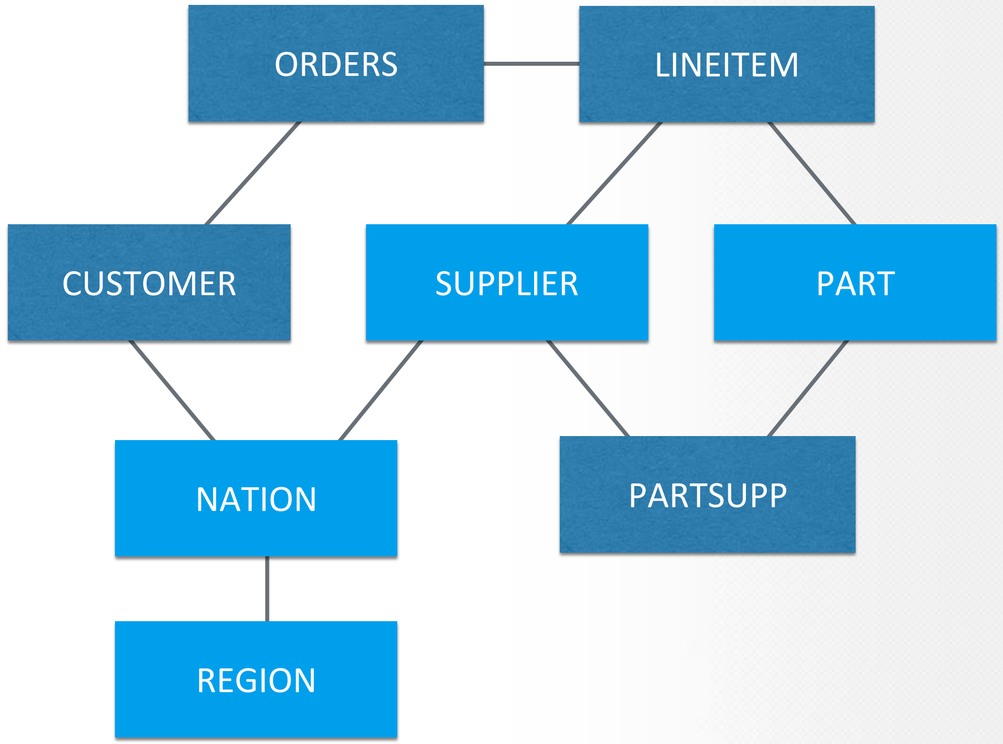
解决了Kylin 1.0很多功能限制：

- 从星形模型到雪花模型
- 单表重复Join
- ……

TPC-H Benchmarking

TPC-H is a benchmark for decision support system.

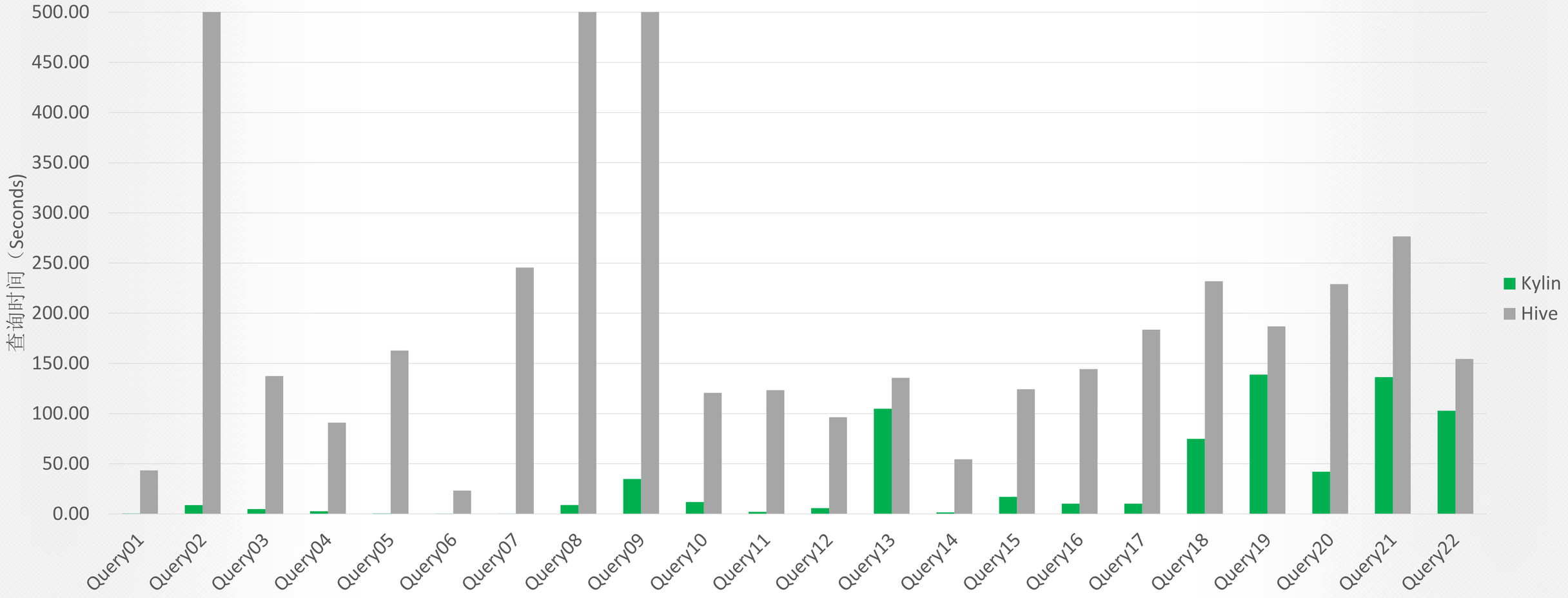
- Popular among commercial RDBMS & DW solutions
- Queries and data have broad industry-wide relevance
- Examine large volumes of data
- Execute queries with a high degree of complexity
- Give answers to critical business questions



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

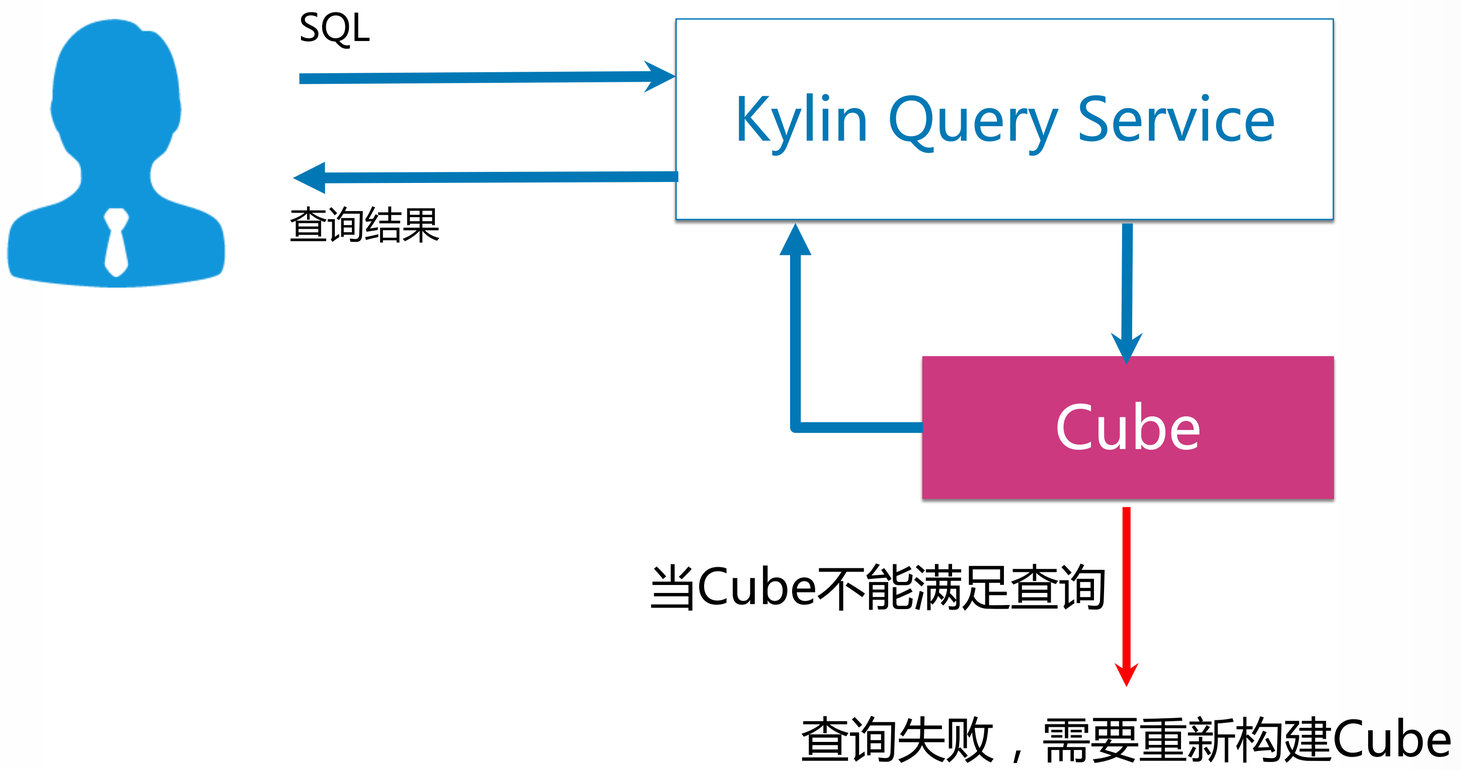
Kylin vs Hive



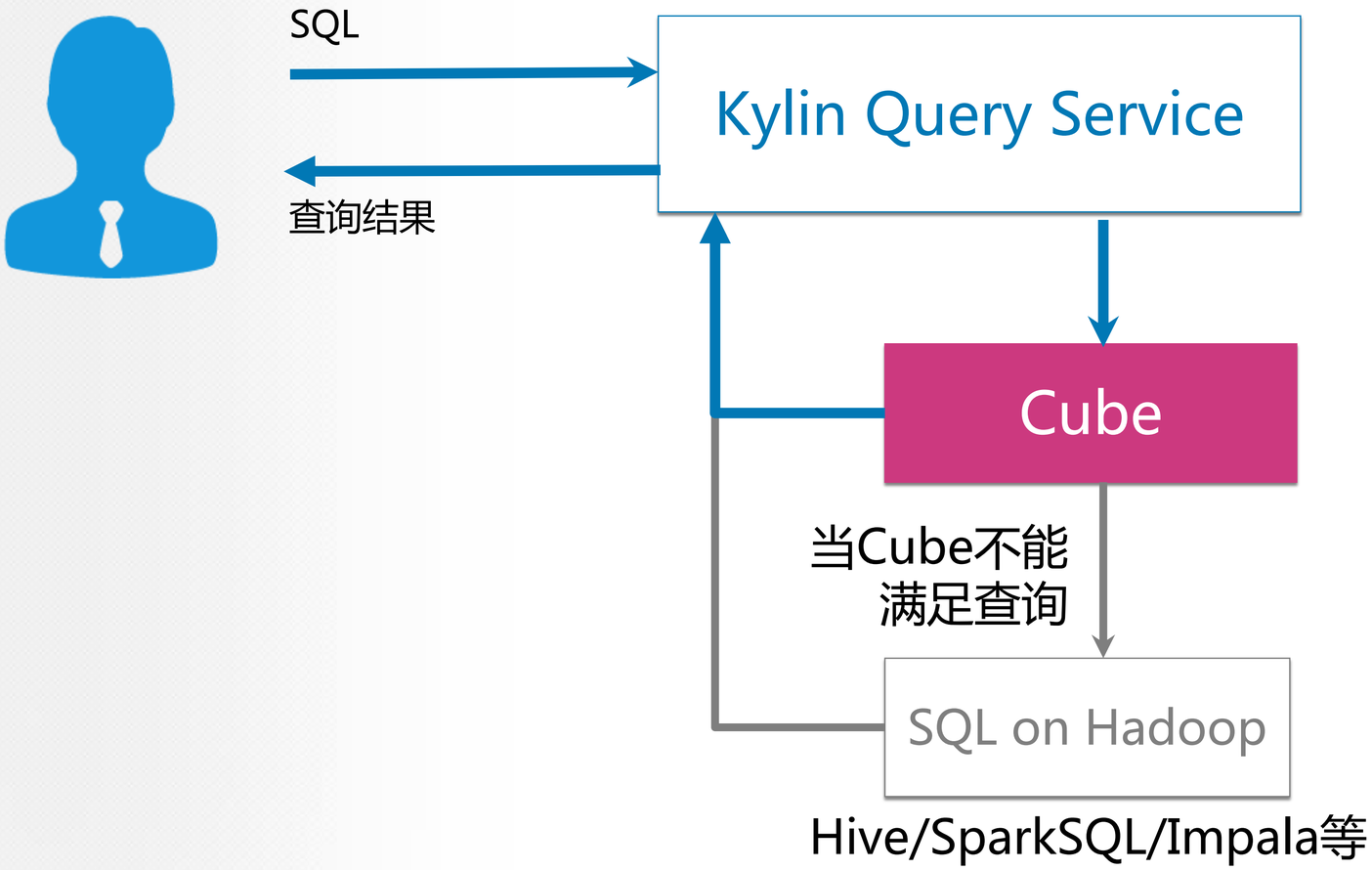
OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识分享平台

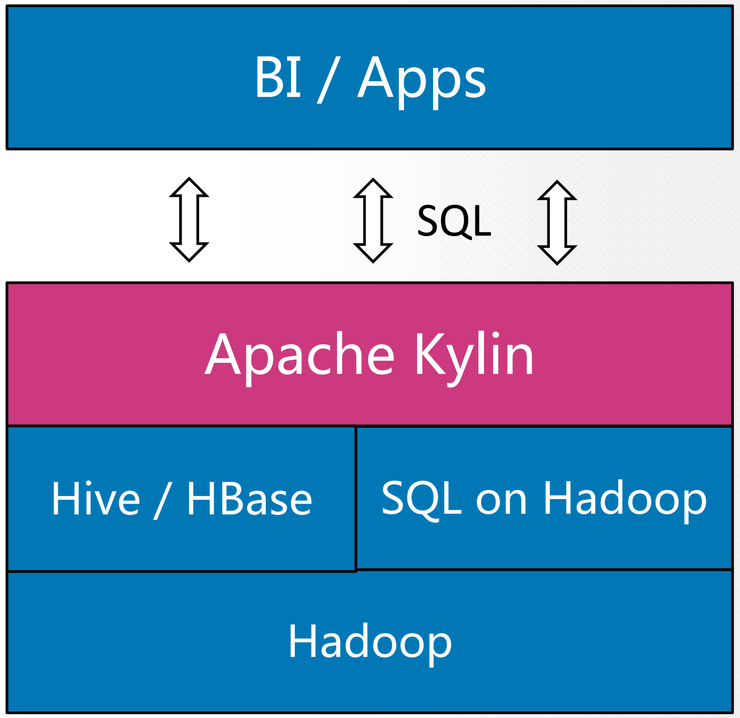
SQL Pushdown

v1.0 - 当Cube不能满足查询



v2.1 - 当Cube不能满足查询



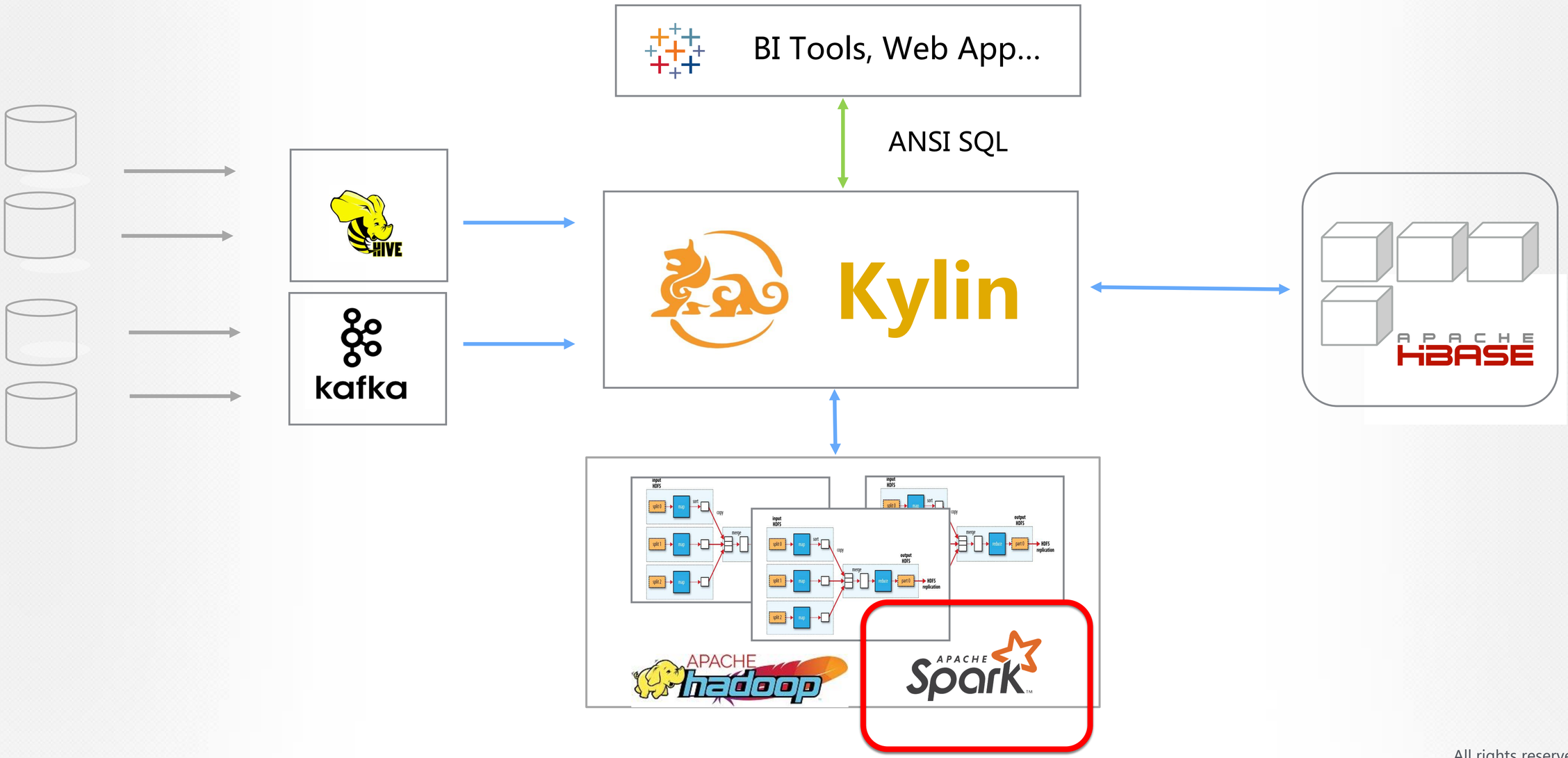


OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识分享平台

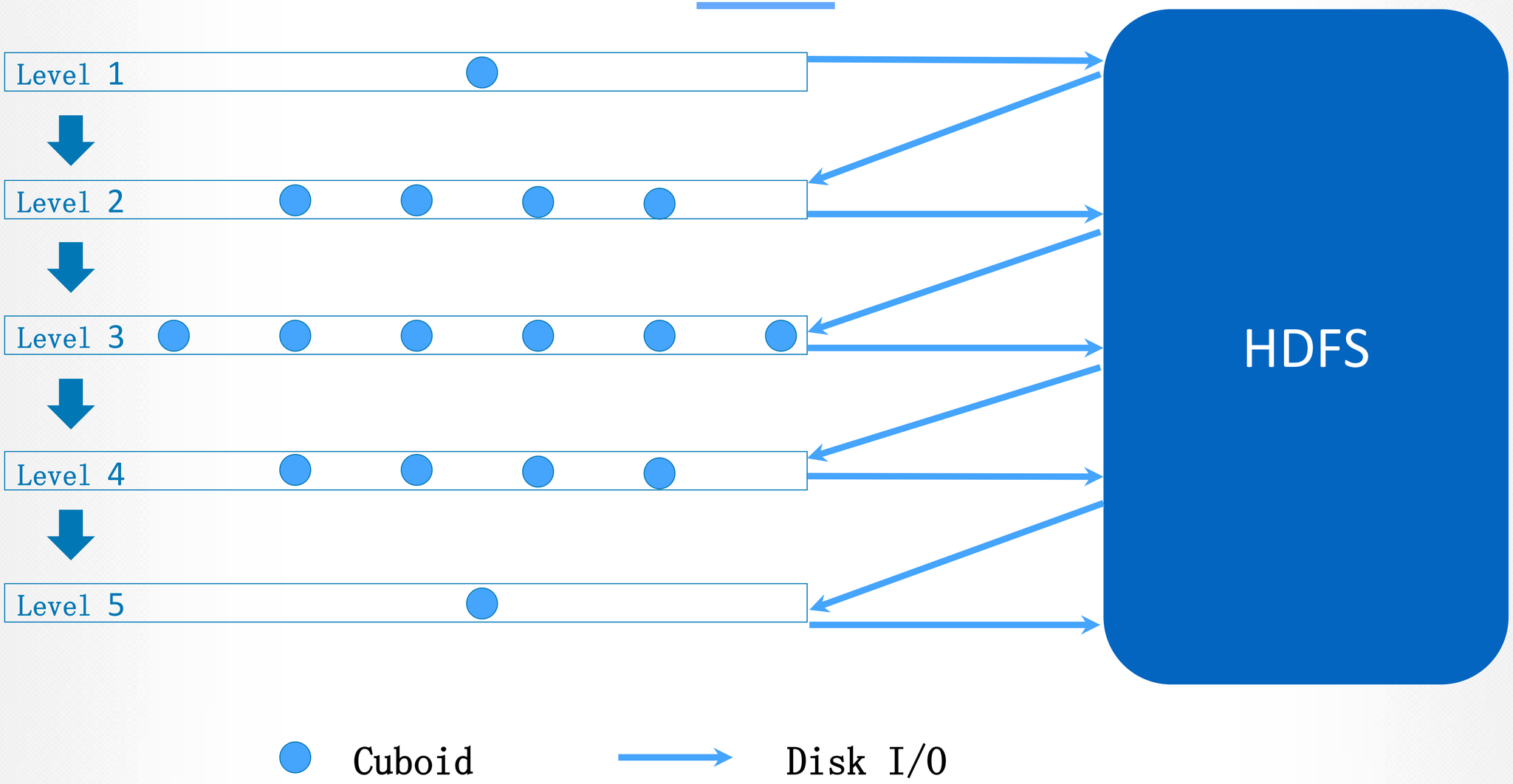
Spark Cubing

Cubing with Spark

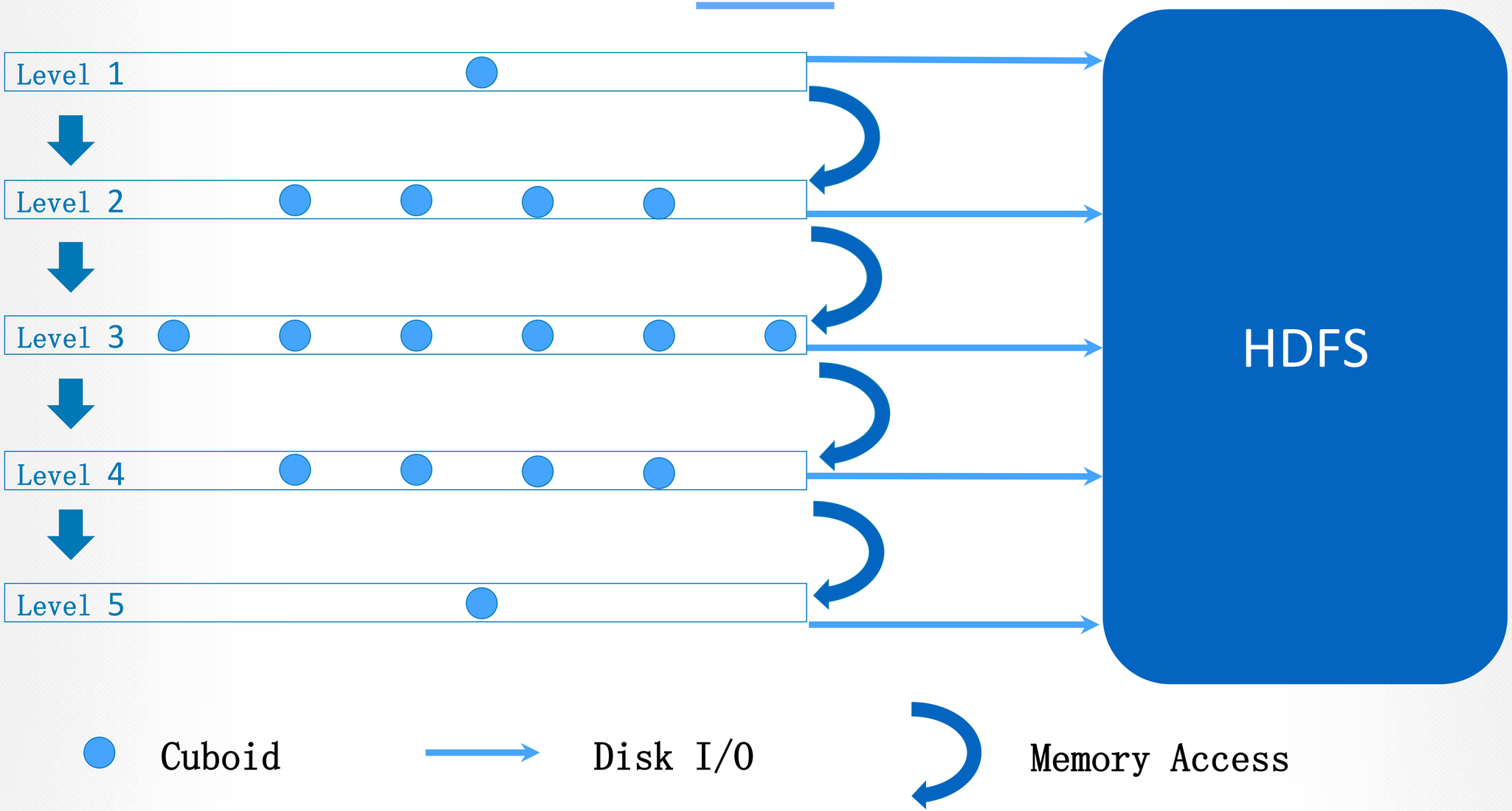
OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识分享平台



MR Layered Cubing



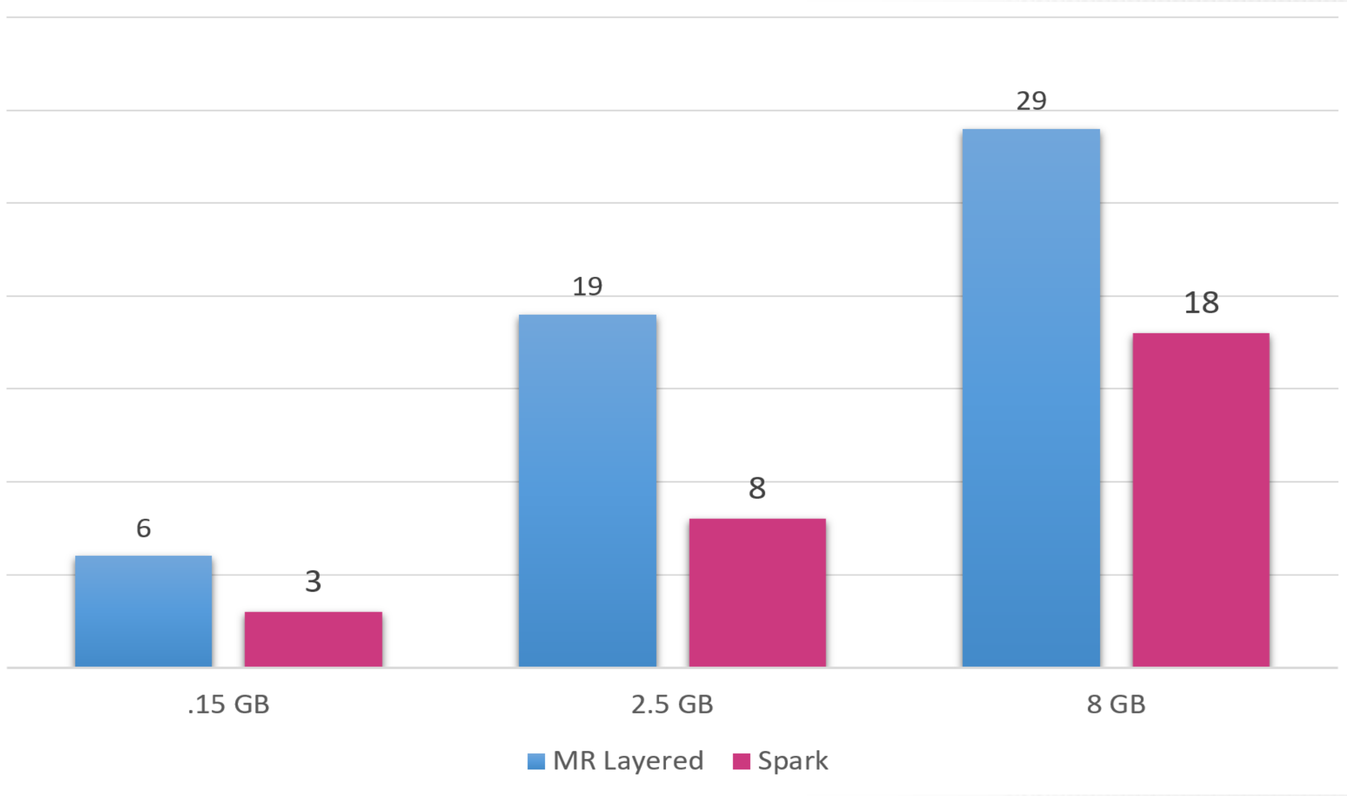
Spark Cubing



Spark Cubing vs. MR Layered Cubing

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

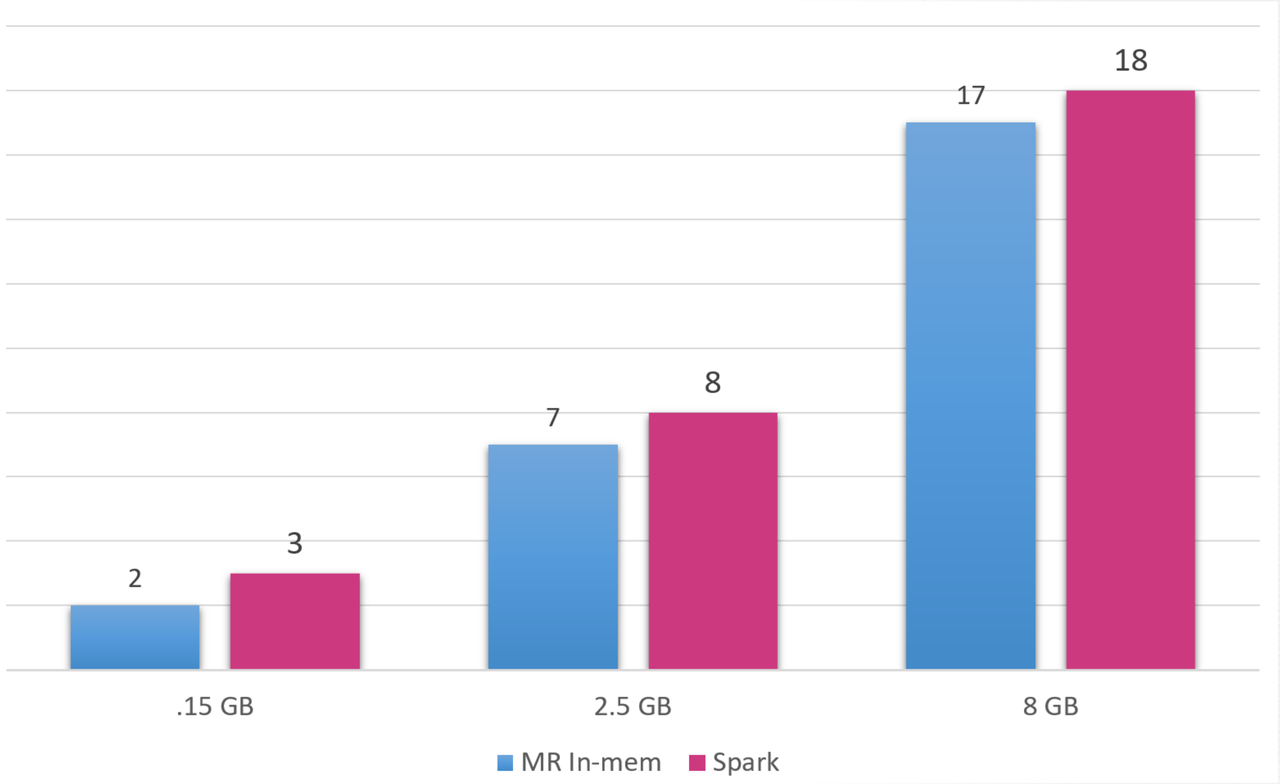


Spark Cubing vs. MR In-mem Cubing

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

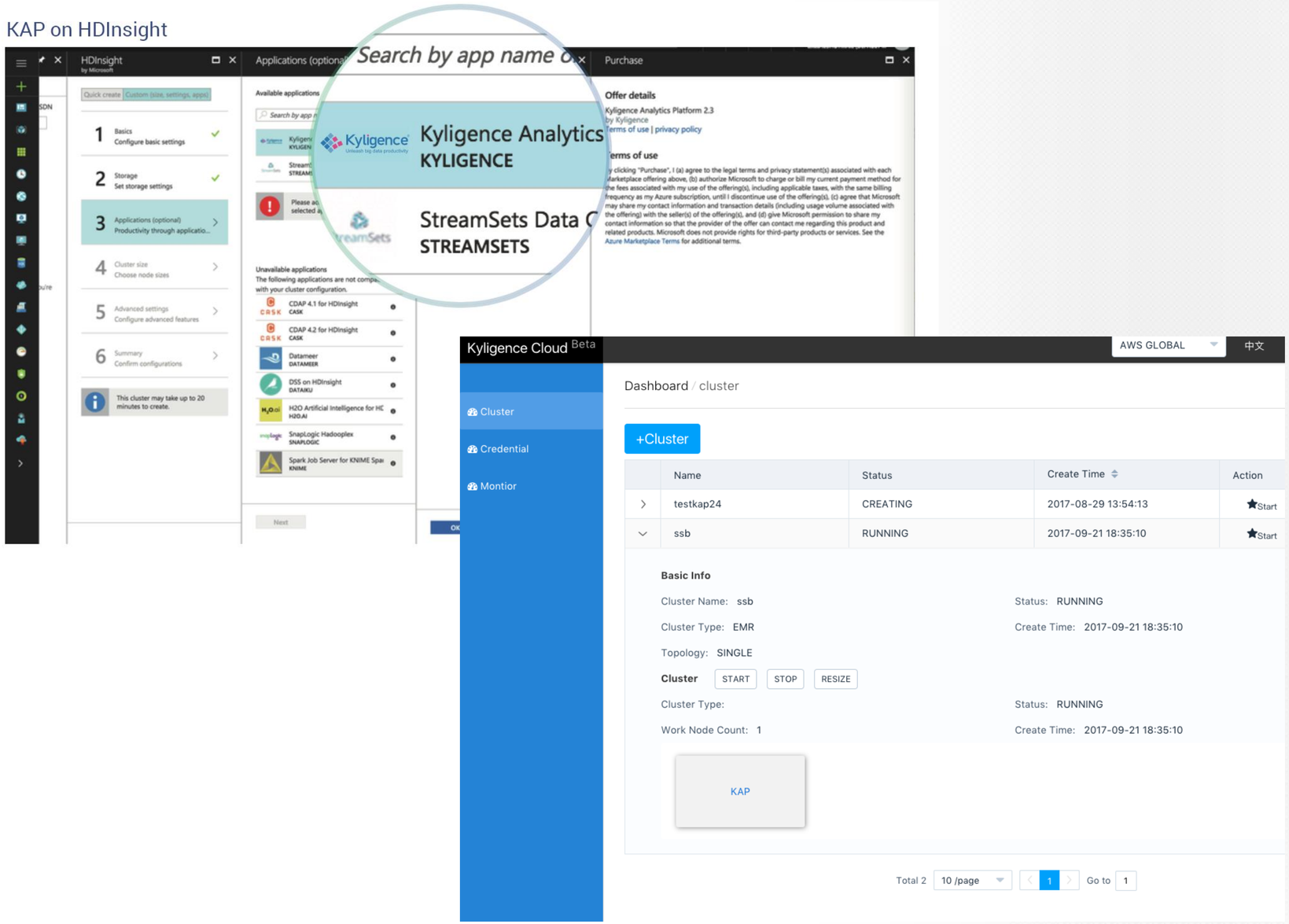
Azure HDInsight

AWS EMR

……

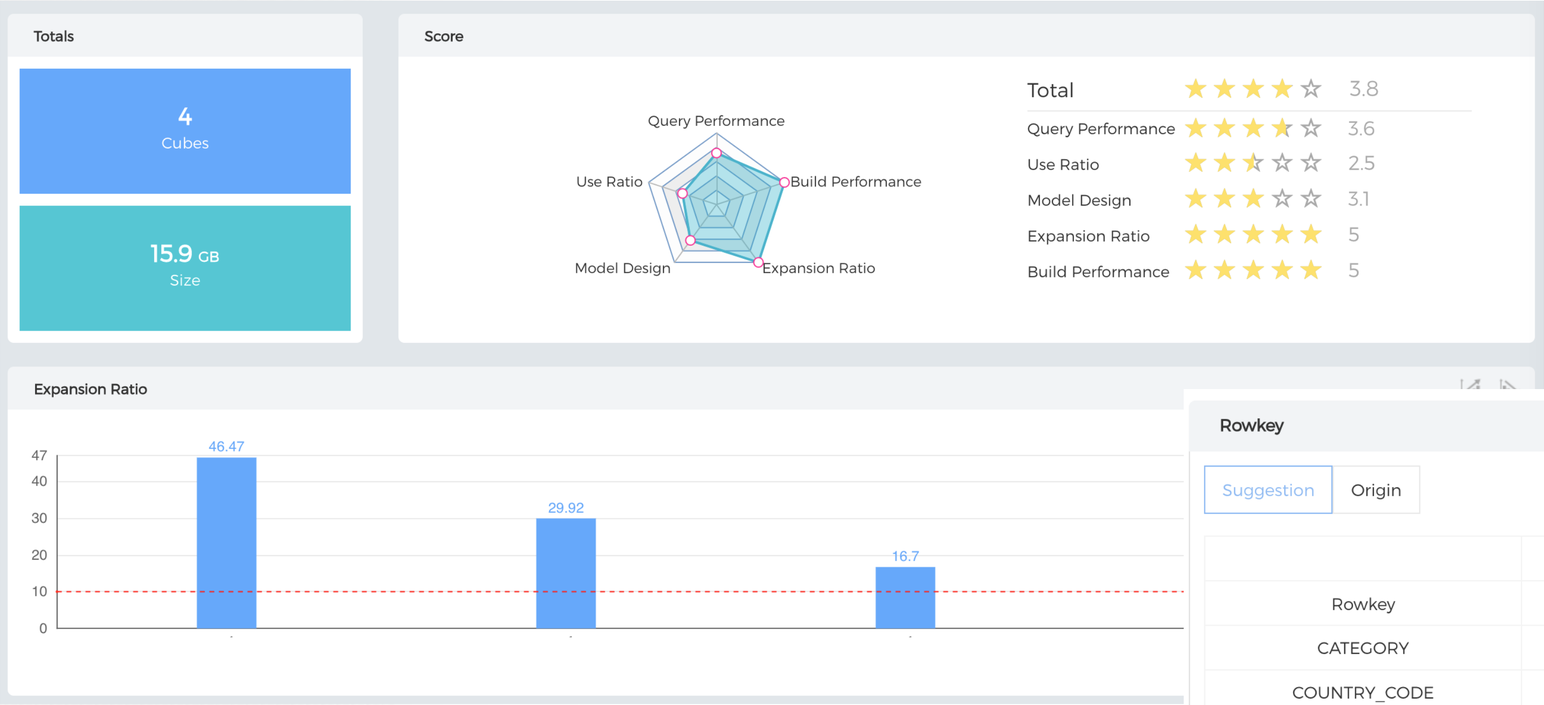
Kyligence Cloud !

KAP on HDInsight



自助式 Cube 调优

OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识分享平台



- 统计评分
- 优化建议

Rowkey

Suggestion | Origin

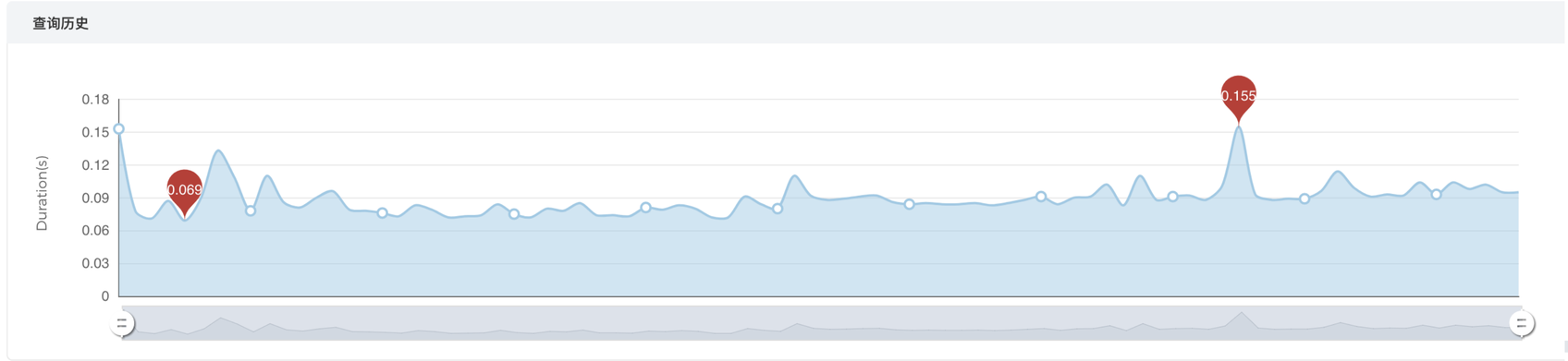
Appear | Absent | difference with parent cuboid

	Basic Information		
Rowkey	Cardinality	Column Type	Encoding
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

Kyligence Robot !

https://kybot.io

自助式查询优化



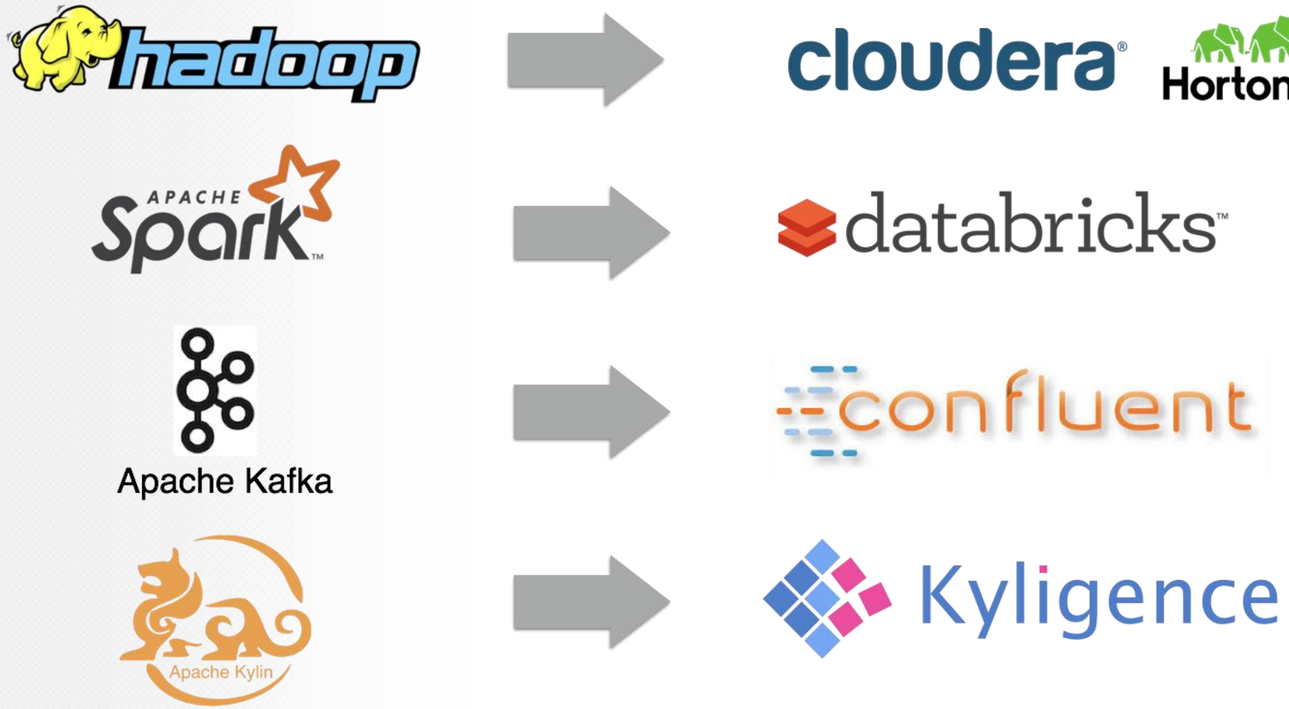
- 历史统计
- 瓶颈分析

Kyligence Robot !

https://kybot.io

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

OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识分享平台

Apache Kylin v2.2 is released!

http://kylin.apache.org

We are hiring!

OSC 源创会 Opensource Innovation Meetup | IT大咖说 知识分享平台

谢谢!

李栋