HUAWEI

Day1-2 MapReduce服务
课程资料

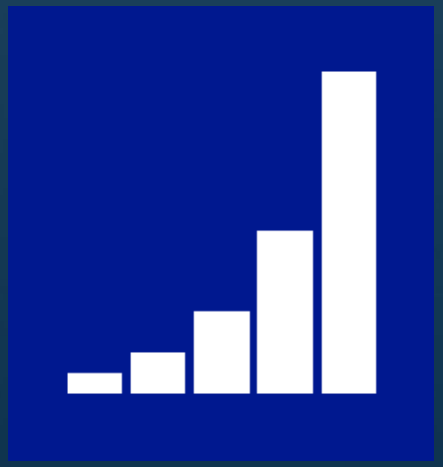
详细视频讲解网址
https://education.huaweicloud.com:8443/courses/course-v1:HuaweiX+CBUCNXE006+Self-paced/about?isAuth=0&cfrom=hwc

目录

大数据介绍　　MRS服务介绍　　动手实践

HUAWEI

大数据带来的挑战



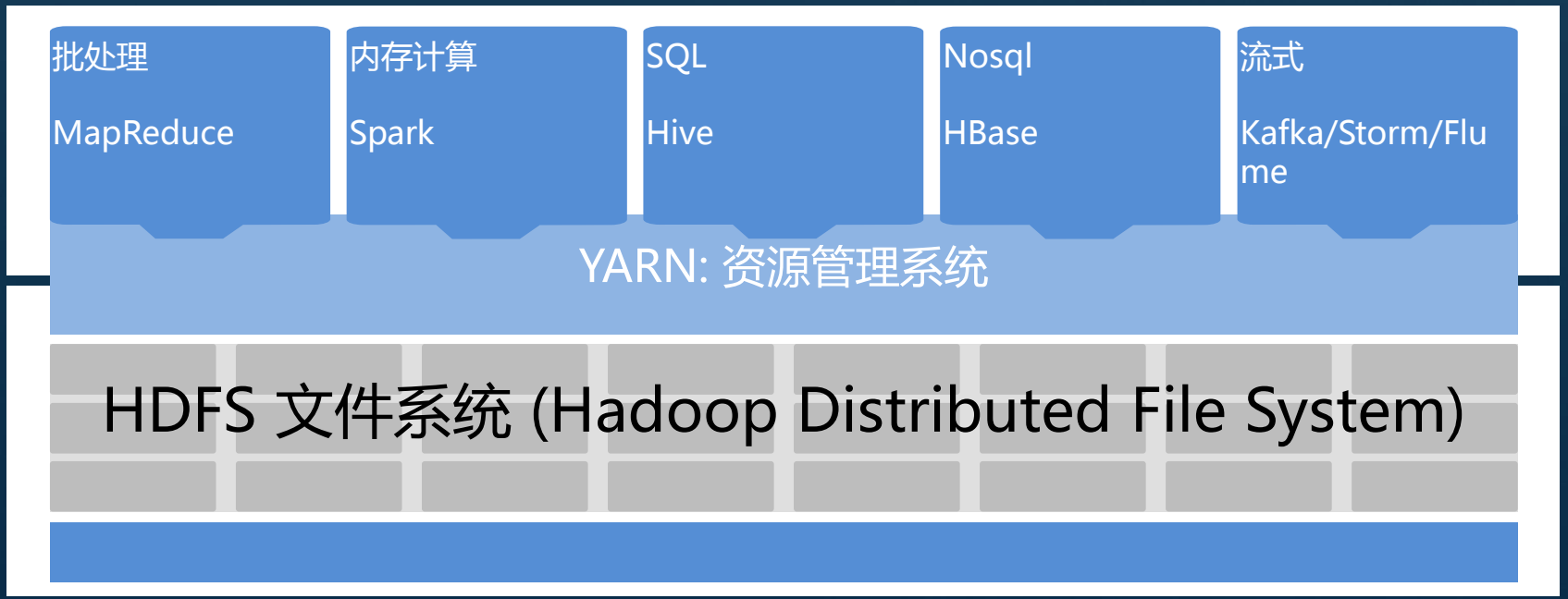
数据量越来越大

数据种类越来越多

数据产生的速度越来越快

Hadoop:大数据的开源解决方案

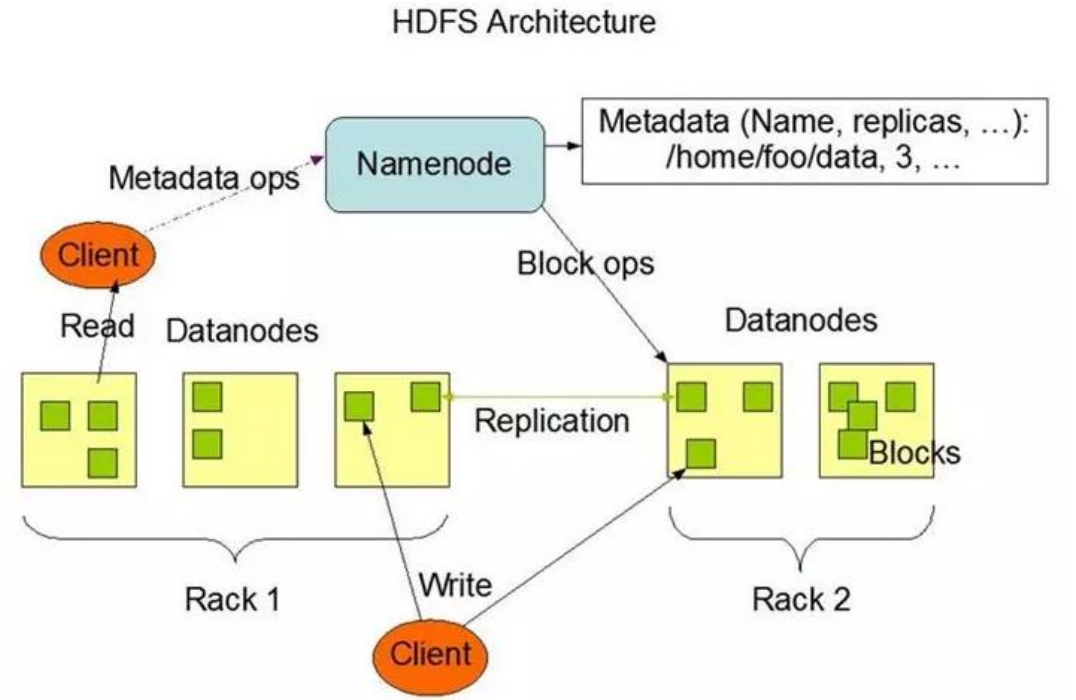
- Apache开源项目，起始于2005年
- 针对解决数据量大，种类多，产生速度快的问题
- 强大的开源社区支持
- 日益丰富的生态系统



Hadoop:大数据的开源解决方案

HDFS

- HDFS是基于Google发布的GFS论文进行设计开发，运行在通用硬件上的分布式文件系统。
- HDFS的特点：
  - 高容错性：认为硬件总是不可靠的，所以每份数据都有备份文件。
  - 高吞吐量：为大量数据访问的应用提供高吞吐量支持
  - 大文件存储：支持存储TB-PB级别的数据



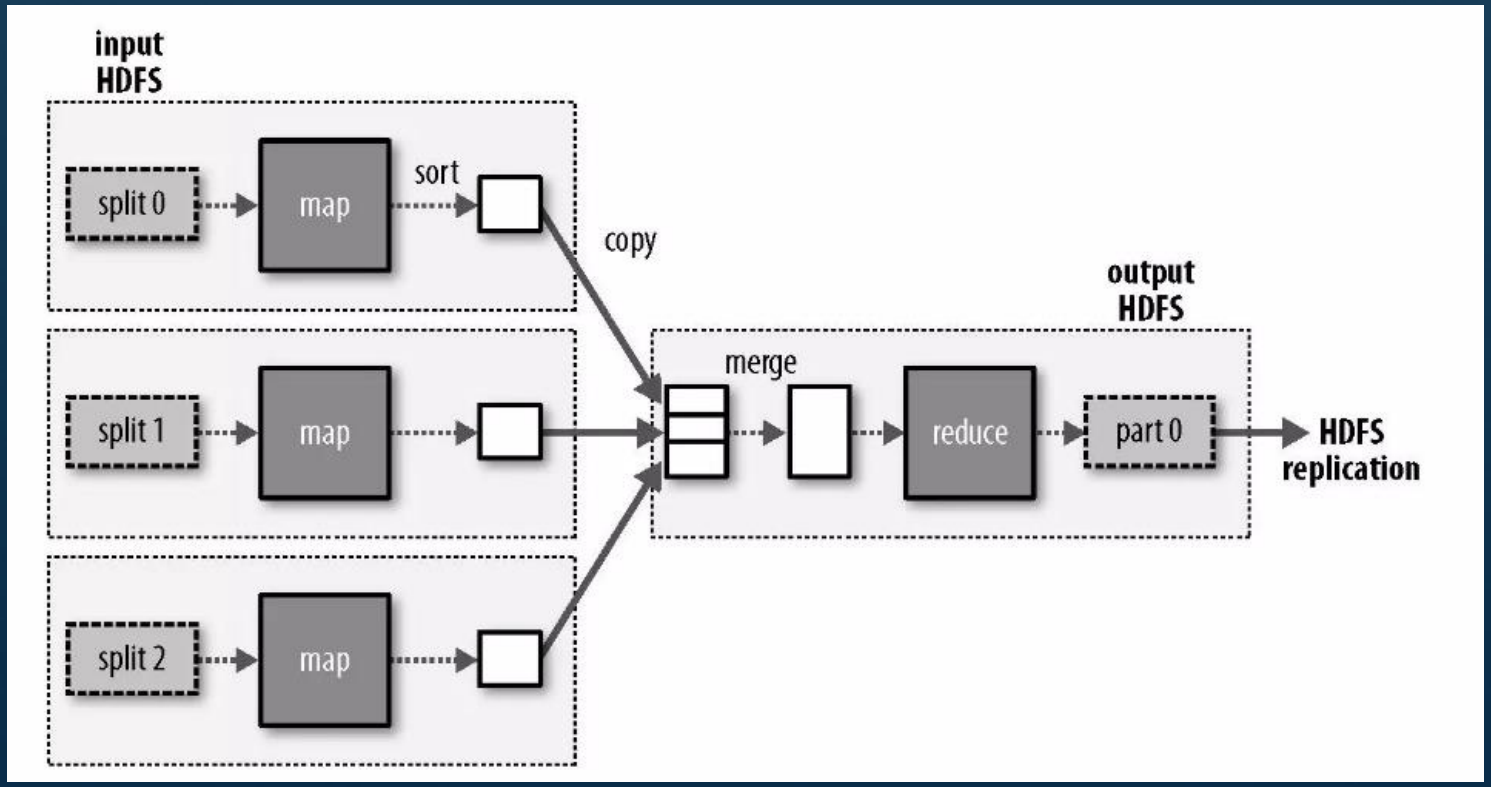
HDFS常用的三个概念：

- NameNode：NameNode用于存储、生成文件系统的元数据信息。
- DataNode：DataNode用于存储实际的数据，将自己管理的数据信息上报给NameNode 。
- Client：支持业务访问HDFS，从NameNode ，DataNode获取数据返回给业务。

Hadoop:大数据的开源解决方案

MapReduce

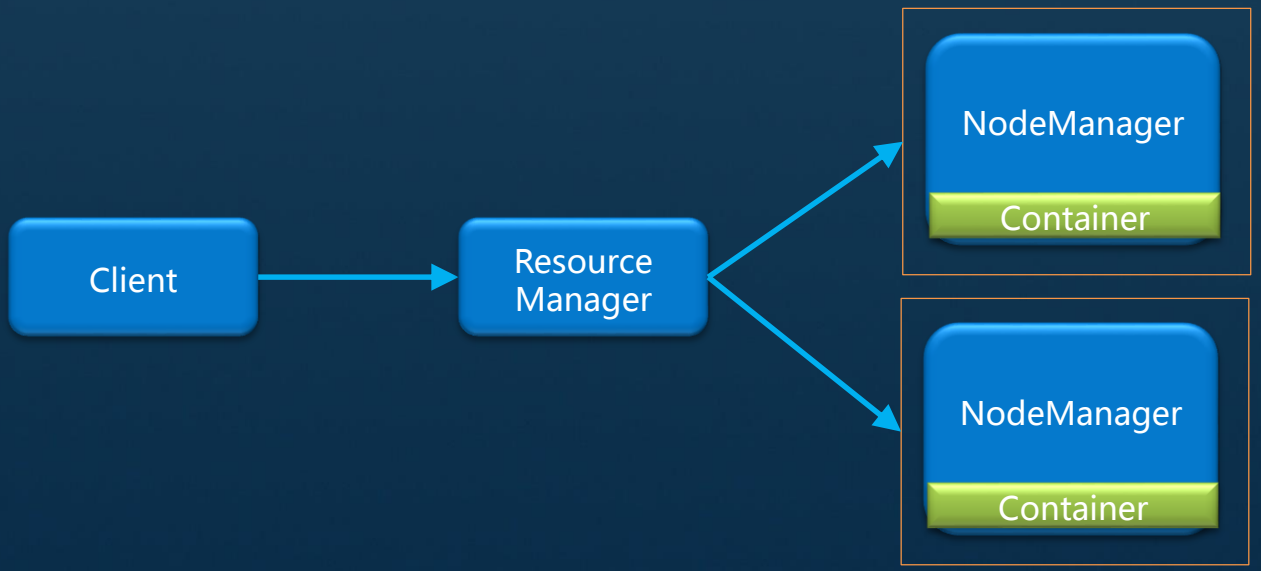
- MapReduce基于Google发布的分布式计算框架Map/Reduce论文设计开发，用于大规模数据集的并行运算，特点如下：
  - 易于编程：程序员仅需描述做什么，具体怎么做交由系统的执行框架处理。



Hadoop:大数据的开源解决方案

Yarn

- Yarn是Hadoop2.0中的资源管理系统，它是一个通用的资源管理模块，可为各类应用程序进行资源管理和作业调度，除了提供MapReduce框架，还可以支持其他框架，比如Spark、Storm等，特点如下：
  - 良好的扩展性：可通过添加节点以扩展集群能力。
  - 高容错性：通过计算迁移策略提高集群的容错性。



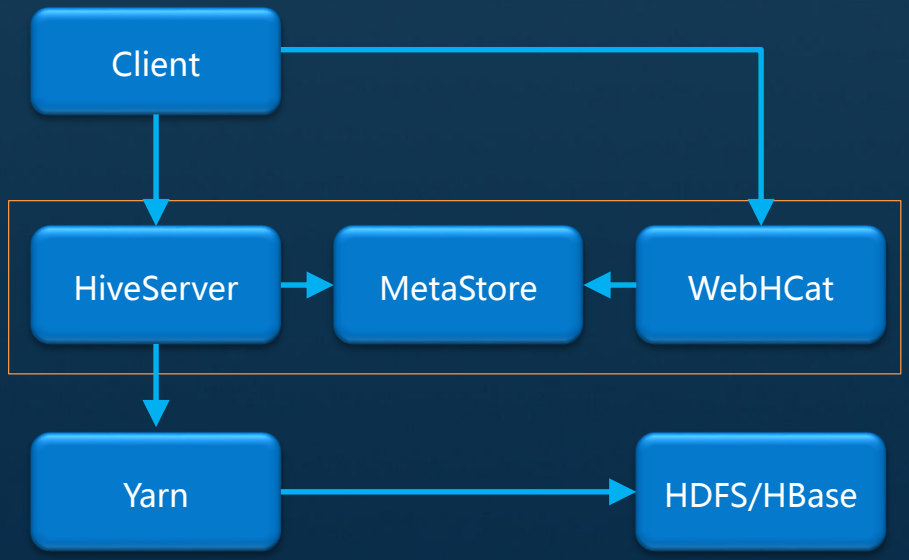
Hadoop:大数据的开源解决方案

Hive

- Hive是基于Hadoop的数据仓库软件，可以查询和管理PB级别的分布式数据。提供类SQL的HiveQL 语言将SQL查询转换为MapReduce任务实现数据处理。

Hive常见场景

- 数据清洗：数据抽取、数据加载、数据转换
- 非实时分析：日志分析、文本分析等
- 数据挖掘：用户行为分析、兴趣分区等



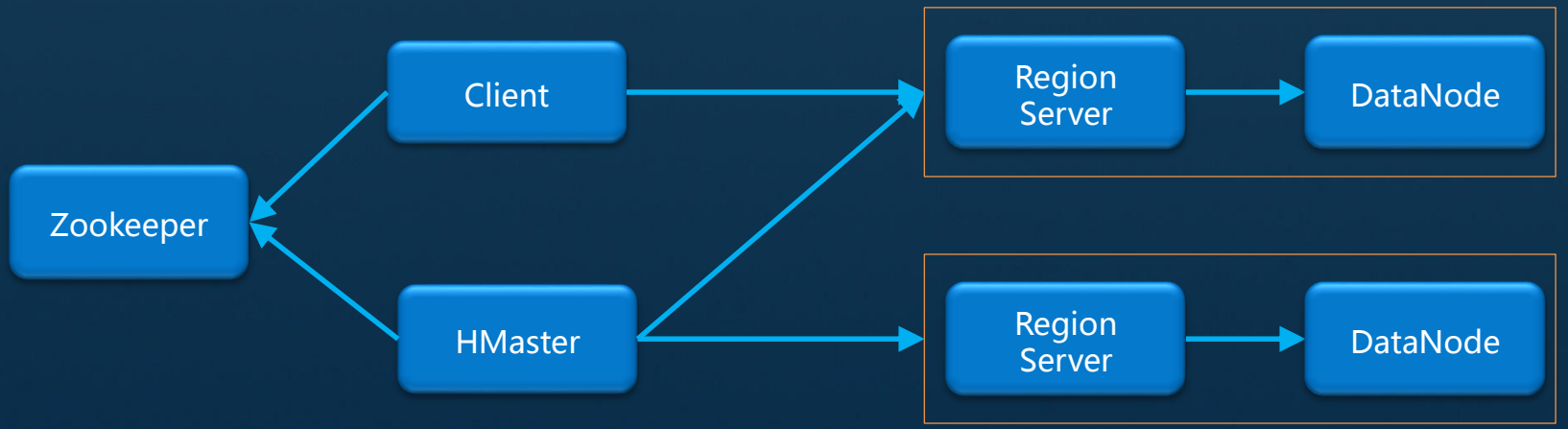
Hadoop:大数据的开源解决方案

HBase

- HBase是一个高可靠性、高性能、面向列、可伸缩的分布式数据库，提供海量数据存储功能，用来解决关系型数据库在处理海量数据时的局限性。

HBase常见场景

- 存储大表数据（表的规模可以达到数十亿行以及数百万列）
- 高效的随机读取
- 同时处理结构化和非结构化的数据



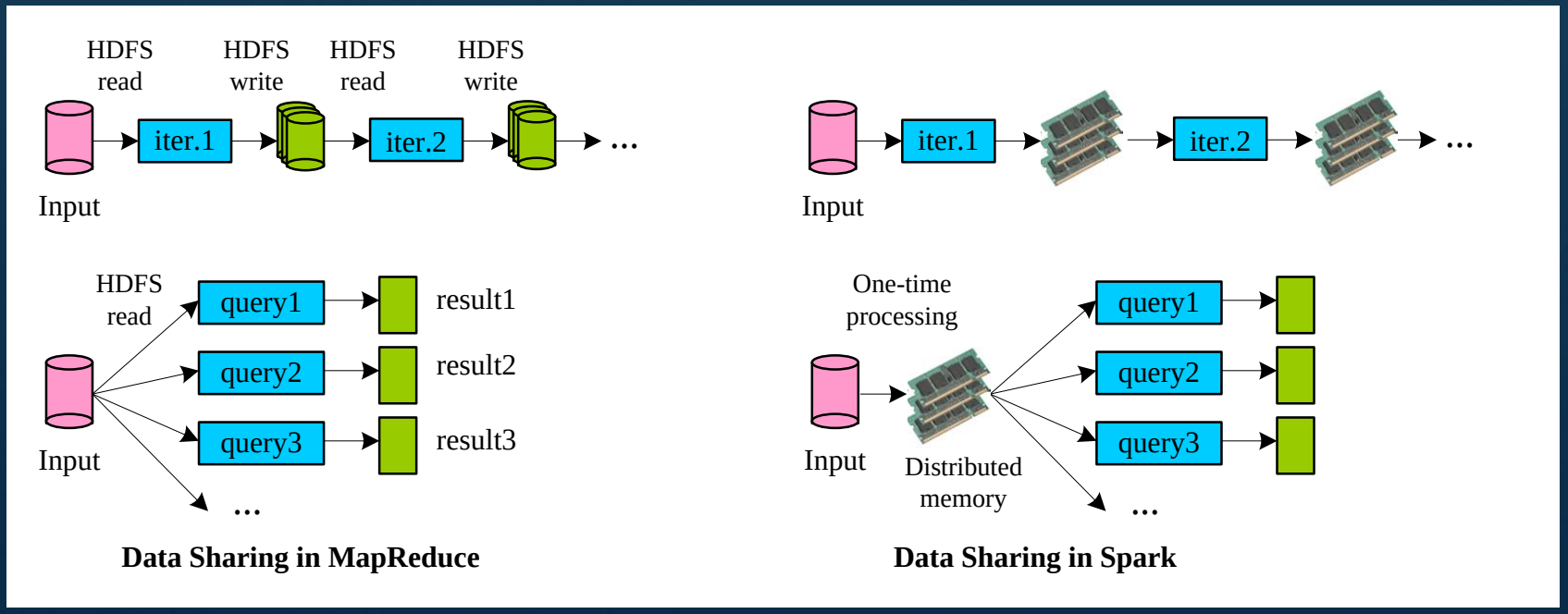
Hadoop:大数据的开源解决方案

Spark

- Spark是一种通用的高性能集群计算系统。既有类似于MR的分布式内存计算框架，也有类似Hive的类SQL查询，还提供了实时数据的处理引擎和机器学习的算法库。

Spark常见场景

- 快速的数据处理，ETL（抽取、转换、加载）
- 实时数据分析
- 数据挖掘和机器学习



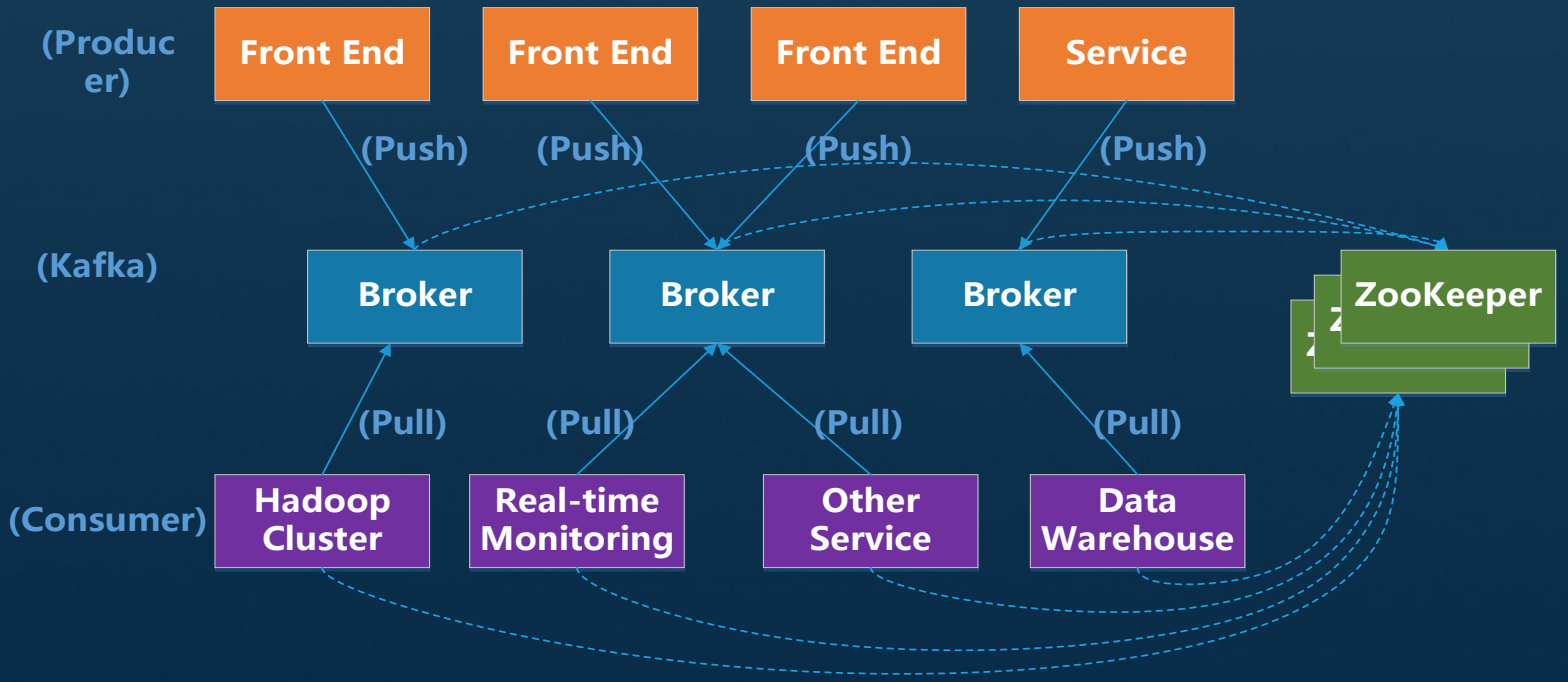
Hadoop:大数据的开源解决方案

Kafka

- Kafka是一个高吞吐、分布式、基于发布订阅的消息系统，利用Kafka技术可在廉价的机器上搭建起大规模消息系统，适用于离线和在线的消息消费。

Kafka常见场景

- 常规的消息收集
- 网站活性跟踪
- 聚合统计系统运营数据（如监控数据）



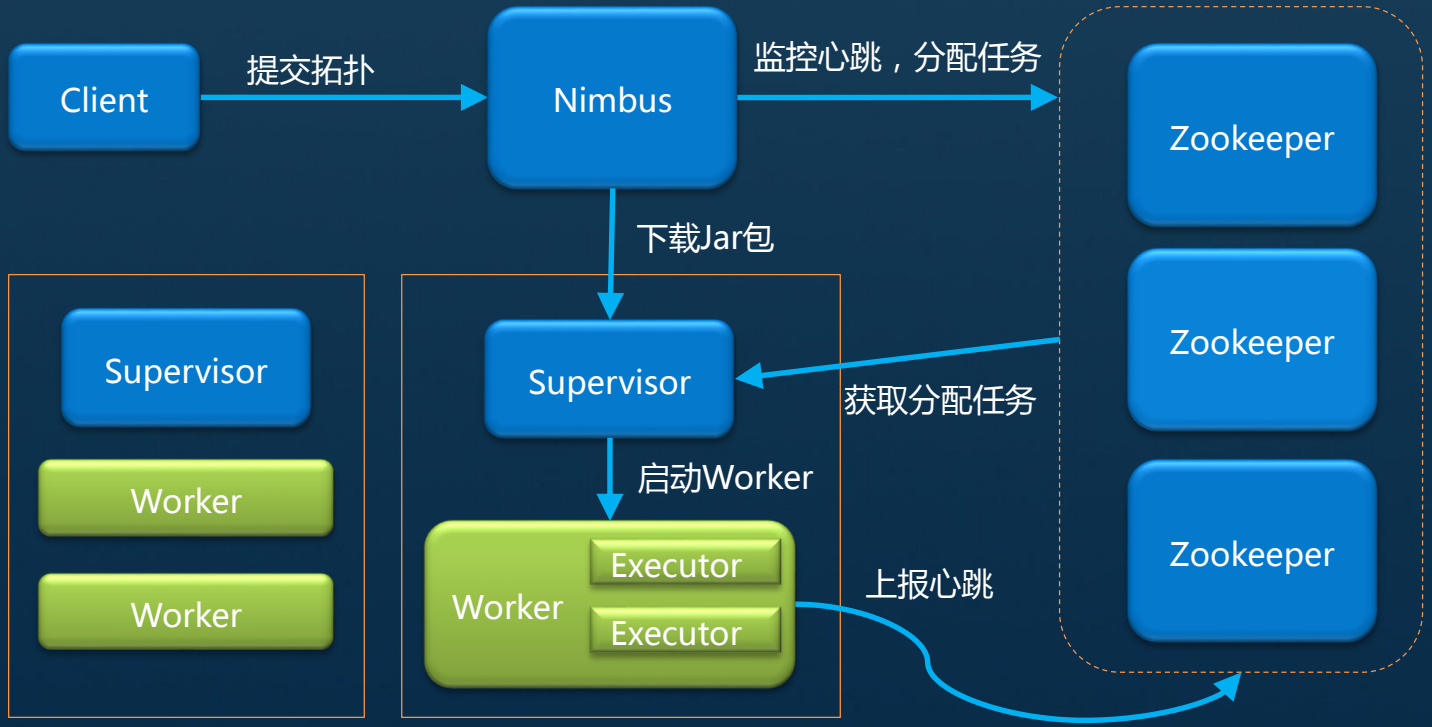
Hadoop:大数据的开源解决方案

Storm

- Storm是一个分布式、实时计算框架，具有高度容错、低时延的优点。

Storm常见场景

- 实时分析：如实时日志处理、交通流量分析等
- 实时统计：如网站的实时访问统计、排序等
- 实时推荐：如实时广告定位、事件营销等



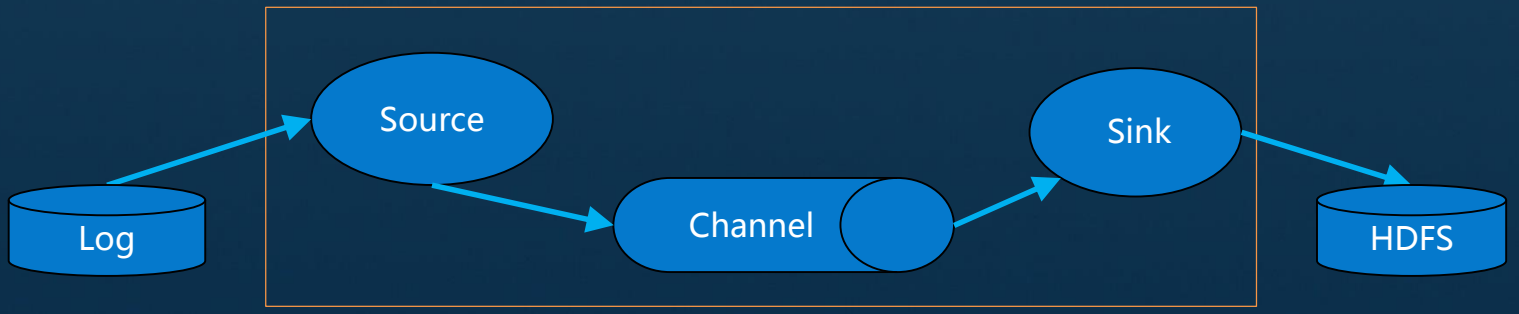
Hadoop:大数据的开源解决方案

Flume

- Flume是一个分布式、可靠和高可用的海量日志聚合的系统。支持在系统中定制各类数据发送方，用于收集数据然后写到各种数据接收方的能力。用户几乎不必进行任何额外开发即可使用。

Flume常见场景

- 从固定目录下采集日志信息到目的地（HDFS, HBase, Kafka)
- 实时采集日志信息到目的地



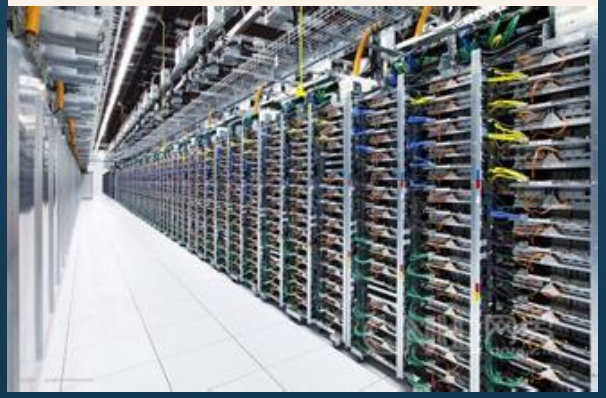
目录

大数据介绍　　MRS服务介绍　　动手实践

HUAWEI

大数据环境搭建痛点

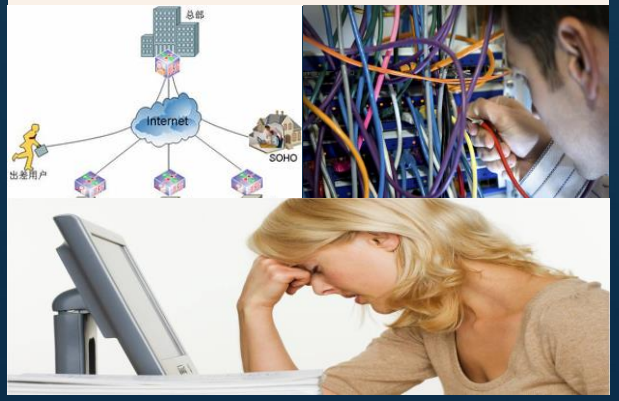
提前建设，建设成本高



业务上线速度慢

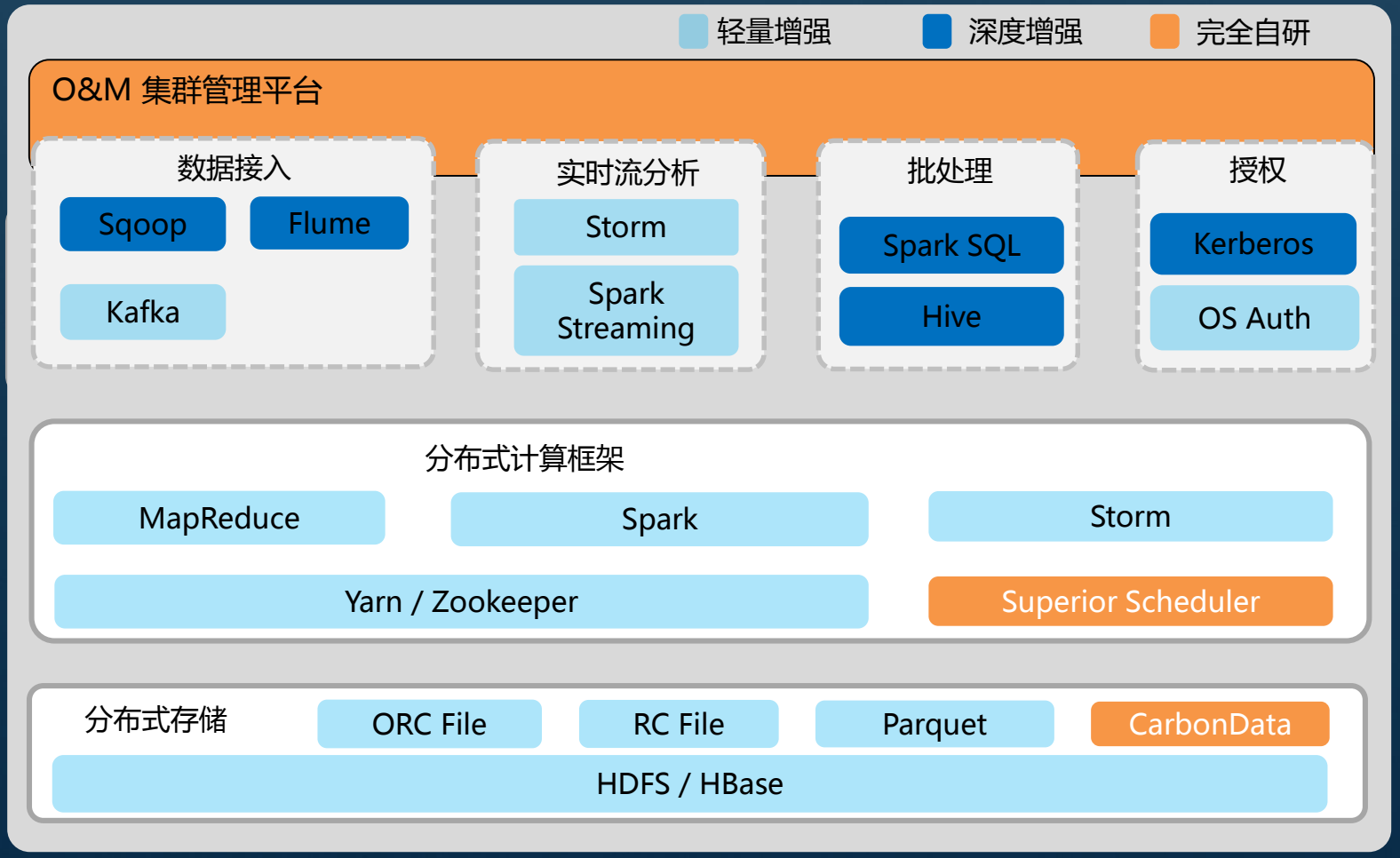
安全性差

维护难度大



HUAWEI

MRS服务——云时代全栈企业级大数据服务

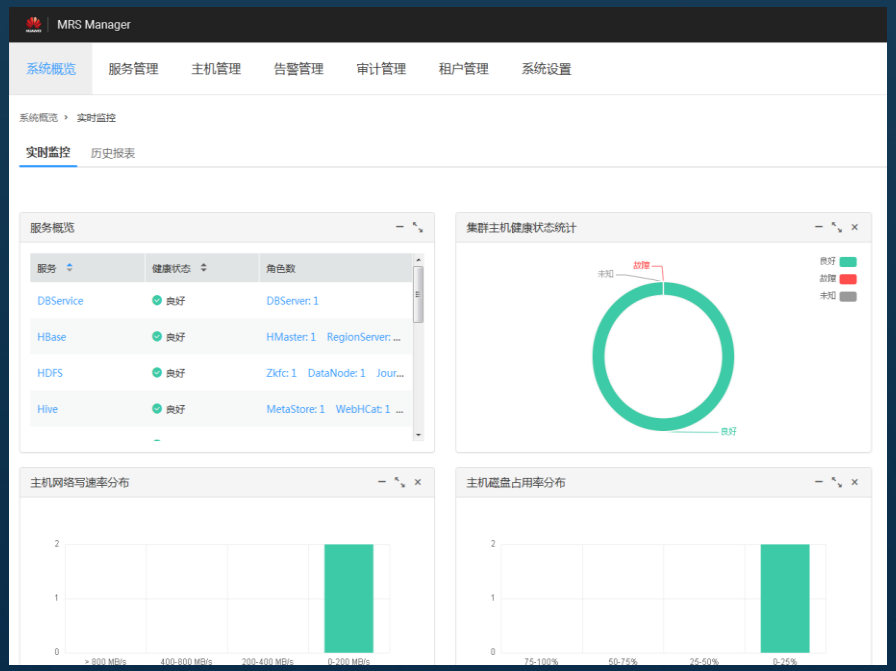


HUAWEI

MRS优势

易运维

- 用户无需关注硬件的购买和维护
- 专门研发的企业级集群管理系统
- 可通过短信/邮件的方式，提醒用户平台异常

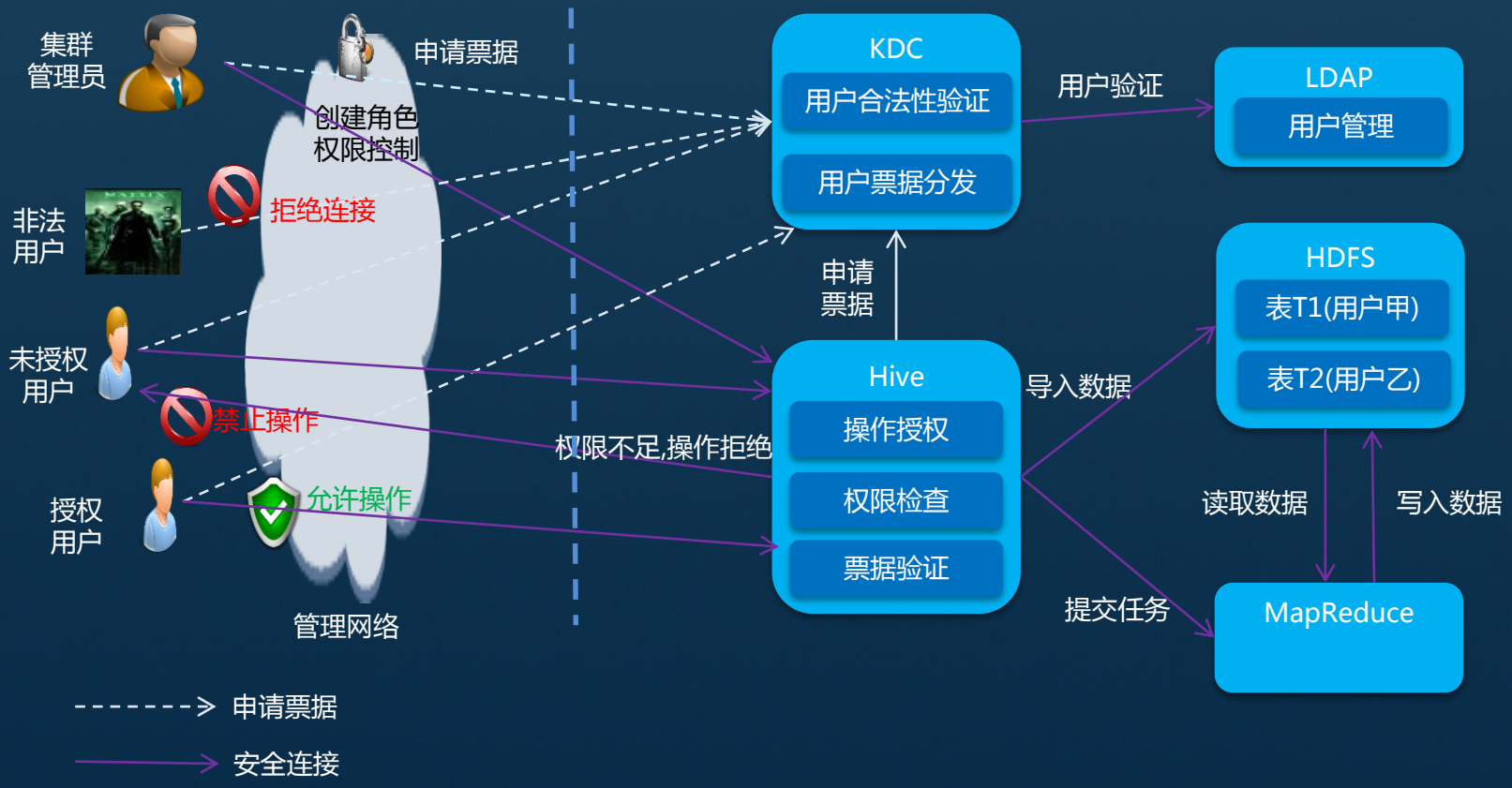


HUAWEI

MRS优势

高安全

- 通过华为专业的安全团队和德国PSA安全认证测试
- Kerberos认证
- 角色级别的安全控制
- 完整的审计功能



HUAWEI

MRS优势

低成本

- 灵活配置节点与磁盘规格
- 支持集群按需扩容、减容
- 支持临时集群，作业运行完自动销毁
- 支持自定义策略，集群自动弹性伸缩

集群名称	配置		价格
mrs_rnMX	集群版本	MRS 1.7.1	¥ 2.25/小时
	集群类型	分析集群	
	地域	华北-北京一	
	项目	缺省	
	计费模式	按需计费	
	可用分区	可用区2	
	虚拟私有云	vpc-mrs-demo	
	子网	subnet-mrs-demo(192.168.0.0/24)	
	安全组	自动创建	
	Master节点规格	c3.xlarge.2 | 4 核 | 8 GB	
	Master节点数量	1	
	Master节点存储空间	100 GB 普通IO * 1	
	Core节点规格	c3.xlarge.2 | 4 核 | 8 GB	
	Core节点数量	1	
	Core节点存储空间	100 GB 普通IO * 1	
	日志记录	关闭	
	Kerberos认证	关闭	
	Hadoop版本	2.8.3	
	Spark版本	2.2.1	
	HBase版本	1.3.1	
	Hive版本	1.2.1	
	Hue版本	3.11.0	
	Loader版本	2.0.0	

配置费用： ¥ 2.25/小时

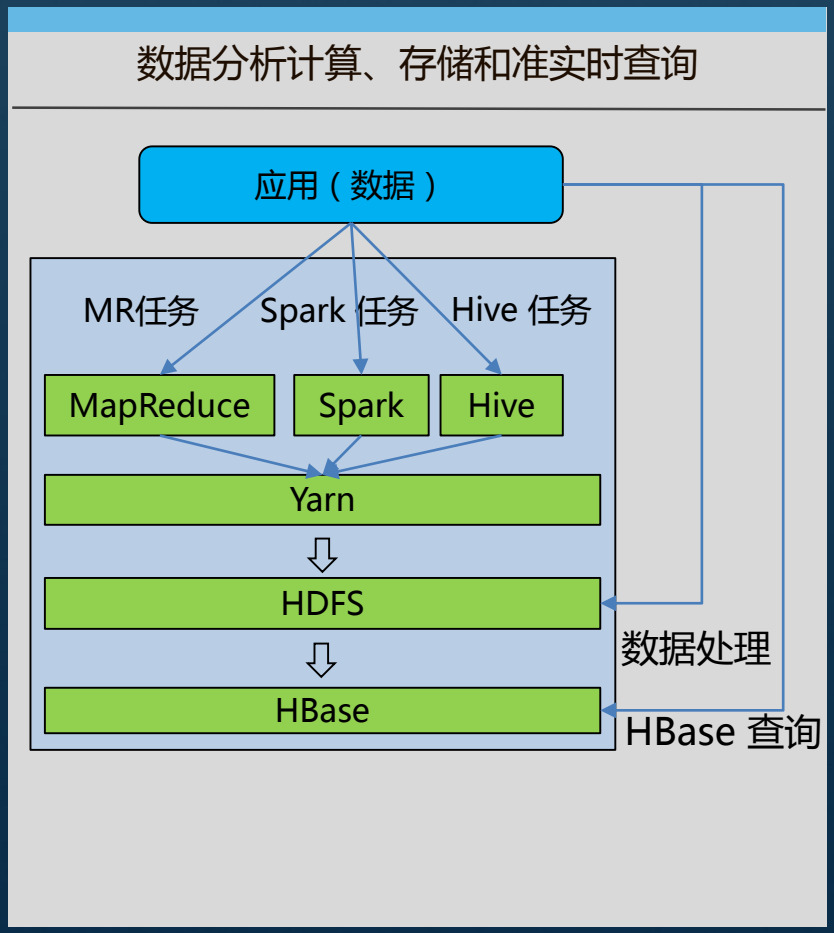
参考价格，具体扣费请以账单为准 了解计费详情

上一步 提交申请

配合华为云生态

围绕数据处理过程，构建全生命周期全栈服务

MRS应用场景



MRS离线分析场景

1.用户将待分析的数据存储到HDFS中

2.用户向MapReduce或Spark或Hive提交数据清洗作业

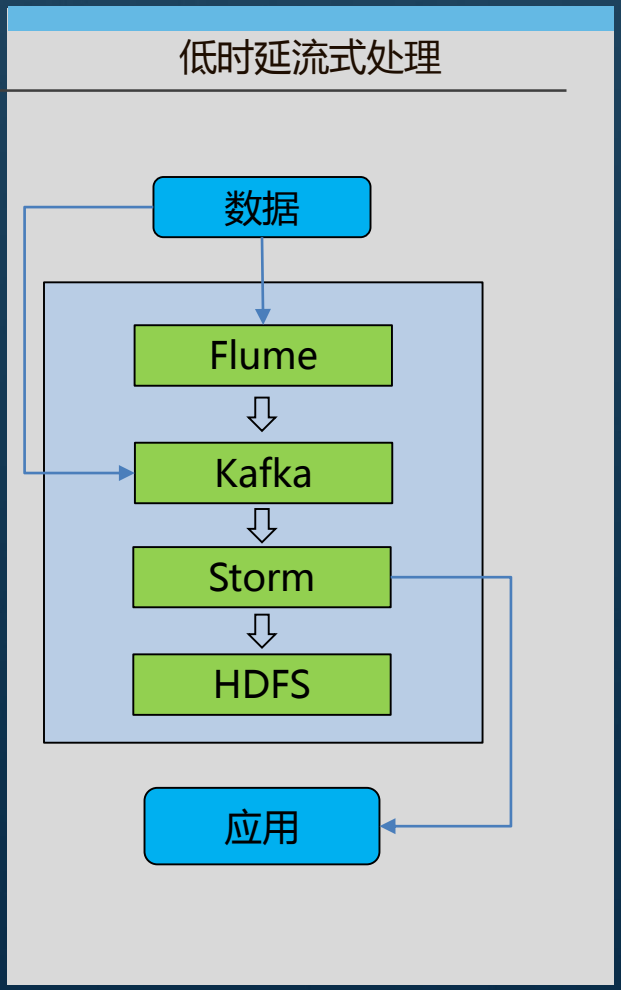
3.在通过Yarn分配资源执行完任务后会将结果数据存放在HDFS中

4.此时用户可以选择将HDFS中的结果数据导入HBase以便查询或直接把结果数据提出到应用做进一步处理

```
likun1000003,华AVM936,32.056444,118.777589,72,211,,2017-01-01 08:00:05,,,,,,,,1,,,
haowei1000008,华A709GB,30.6786,104.070835,143,115,,2017-01-01 08:00:08,,,,,,1,,,,,,
haowei1000008,华A709GB,30.6786,104.070835,120,115,,2017-01-01 08:00:08,,,,,,,1,0,,,,
zouan1000007,华A58M83,28.211549,112.979477,77,177,,2017-01-01 08:00:10
haowei1000008,华A709GB,30.678366,104.071341,99,115,,2017-01-01 08:00:10
```

驾驶员ID	车牌号	急加速次数	急刹车次数	空挡滑行次数	空挡滑行时间总和	超速次数	超速时间总和
hanhui1000002	华AZI419	401	444	327	2844	3349	31813
zouan1000007	华A58M83	360	385	315	2997	3181	31248
haowei1000008	华A709GB	321	314	255	2659	2639	25522
xiezhi1000006	华A6CU11	255	310	254	2074	2535	23942
panxian1000005	华AX542C	395	434	330	2930	3531	33946
shenxian1000004	华ADJ750	374	356	297	2810	3126	31494

MRS应用场景



MRS实时分析场景
1.用户将实时数据使用Flume或者编写Kafka程序导入Kafka中
2.然后编写Storm程序对实时数据进行处理
3.处理后的数据可以存储在HDFS中，也可以提供给应用程序进行展示

```
华AG3852,2018-07-24 05:57:32,Gate0722,121.163516,31.180197
华AB1O6A,2018-07-24 05:57:32,Gate0236,121.212811,31.431303
华A73HG1,2018-07-24 05:57:32,Gate0614,121.252008,31.109444
华AN3HJL,2018-07-24 05:57:32,Gate0477,121.392344,31.149775
华A3X655,2018-07-24 05:57:32,Gate0931,121.420044,31.116667
华A0T3U5,2018-07-24 05:57:32,Gate0868,121.457002,31.160869
华AIIDS0,2018-07-24 05:57:32,Gate0295,121.441616,31.153228
华A6ZZ0Q,2018-07-24 05:57:32,Gate0767,121.308091,31.107614
华AN8907,2018-07-24 05:57:32,Gate0385,121.397963,31.173803
```

目录

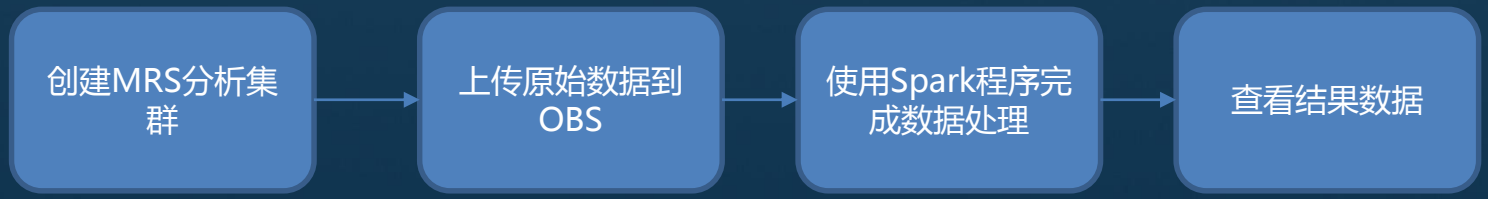
HUAWEI

动手实践

作业一：

下面我们一起来学习如何使用 MRS 来完成离线数据的分析

基本步骤：



实践项目介绍和具体操作步骤指导下载地址：
https://obs-train-data.obs-website.cn-north-1.myhwclouds.com/MRS作业一操作指导书.docx

课后练习

作业二：

本次课后练习需要编写一个MapReduce程序来对驾驶行为数据进行分析，总共分为两个部分：

1. 使用MapReduce程序对违法数据进行统计然后输出和作业一相同结果的文件
2. 使用MapReduce程序对违法数据进行统计并且按照超速次数数据进行升序排序后输出文件，格式与作业一相同

MapReduce二次开发帮助文档：
https://support.huaweicloud.com/devg-mrs/mrs_06_0002.html
https://support.huaweicloud.com/devg-mrs/mrs_06_0082.html

课后练习

参考资料

MRS使用指南：
https://support.huaweicloud.com/usermanual-mrs/mrs_01_0360.html

MRS实践训练营：
https://activity.huaweicloud.com/devcloud_bigdata/index.html?utm_source=guanwang&utm_medium=banner

Thank You.