



场景化大数据应用平台 实践 & 思考

霍秉杰 | 青云QingCloud 大数据平台研发工程师

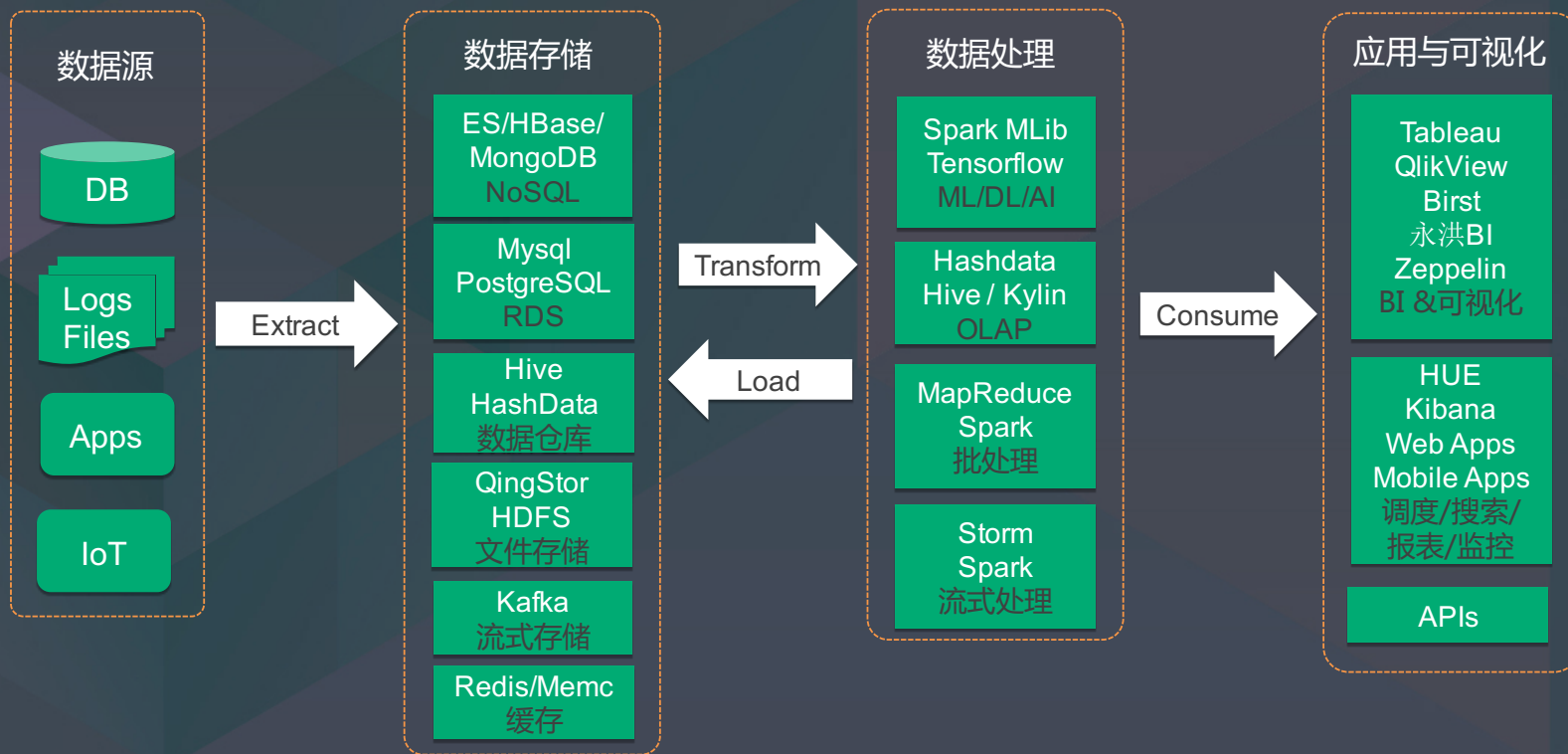
Agenda

- ▶ 大数据分析通用场景
 - ▶ 架构选型
 - ▶ 技术选型
- ▶ 大数据平台与QingStor 对象存储集成场景
 - ▶ ElasticSearch 与 QingStor 对象存储集成
 - ▶ Spark/Hadoop/Hive 与 QingStor 对象存储集成
- ▶ 大数据平台 on AppCenter 2.0
 - ▶ Roadmap
 - ▶ 大数据平台架构
 - ▶ 大数据应用市场
 - ▶ 大数据应用示例

大数据分析通用场景



大数据分析通用场景 - 架构选型



大数据分析通用场景 - 架构选型

	步骤	青云QingCloud 大数据服务
数据采集	Files/logs	Logstash / Flume / qscctl
	Others: JDBC, MongoDB, BlockChain, JMX, Twitter	Canal/Sqoop, Kafka connect
数据存储	File Storage	QingStor 对象存储 / HDFS
	RDS	MySQL/PostgreSQL
	NoSQL	HBase/ElasticSearch/MongoDB
	Stream Storage	Kafka
	Cache	Redis/Memcached
	SQL on Hadoop Data Warehouse	Hive (Kelin, Presto, Impala)
数据处理与分析	SQL on Cloud Data Warehouse	HashData
	Batch Processing	Hadoop/Spark on QingCloud
	Stream Processing	Storm/Spark Streaming
应用与可视化	ML / DL / AI	Spark MLib/Tensorflow
	Consume & BI & Visualization	HUE/Kibana/永洪BI / Zeppelin

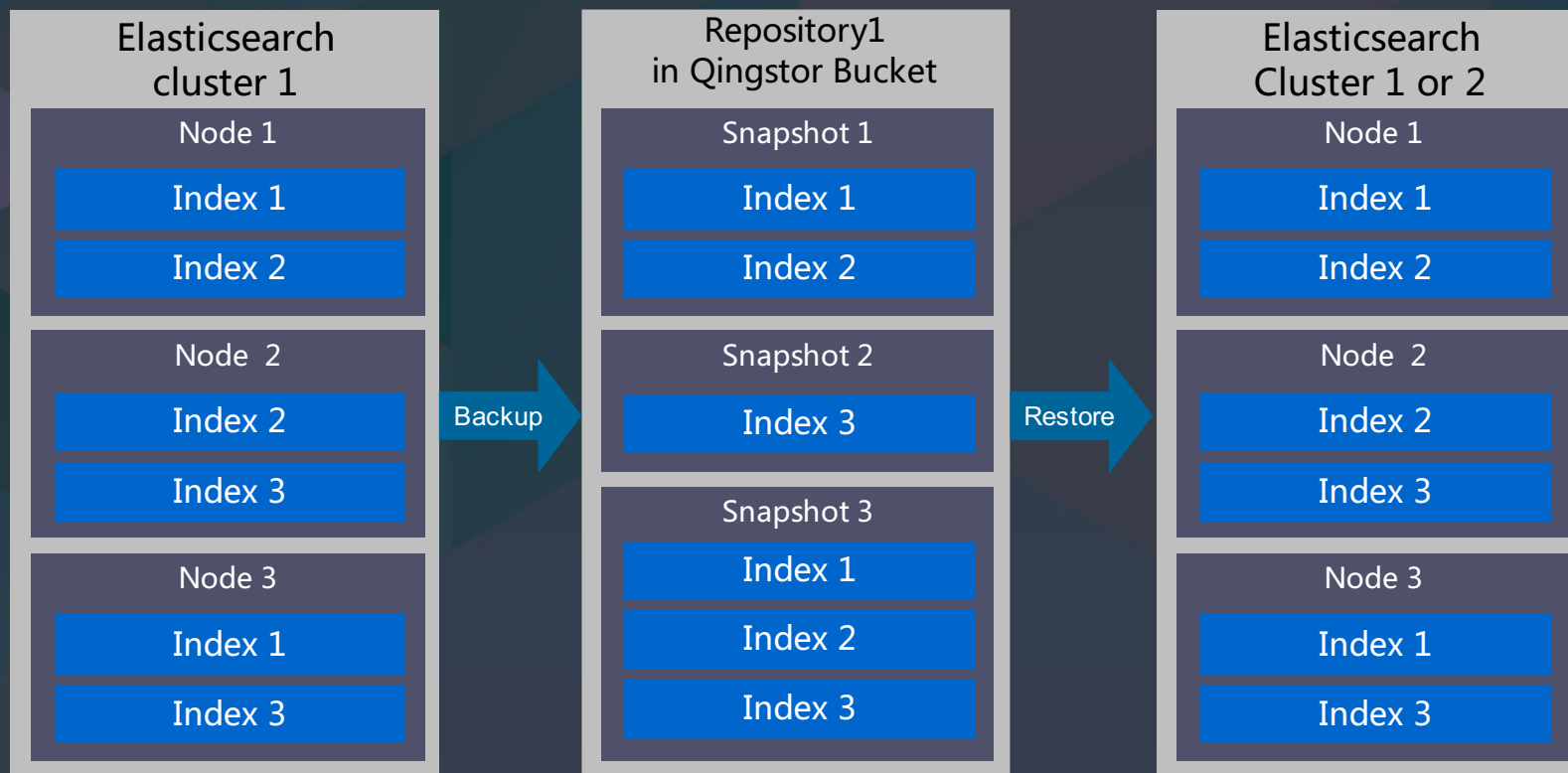


QINGCLOUD 青云

大数据平台与QingStor 对象存储集成



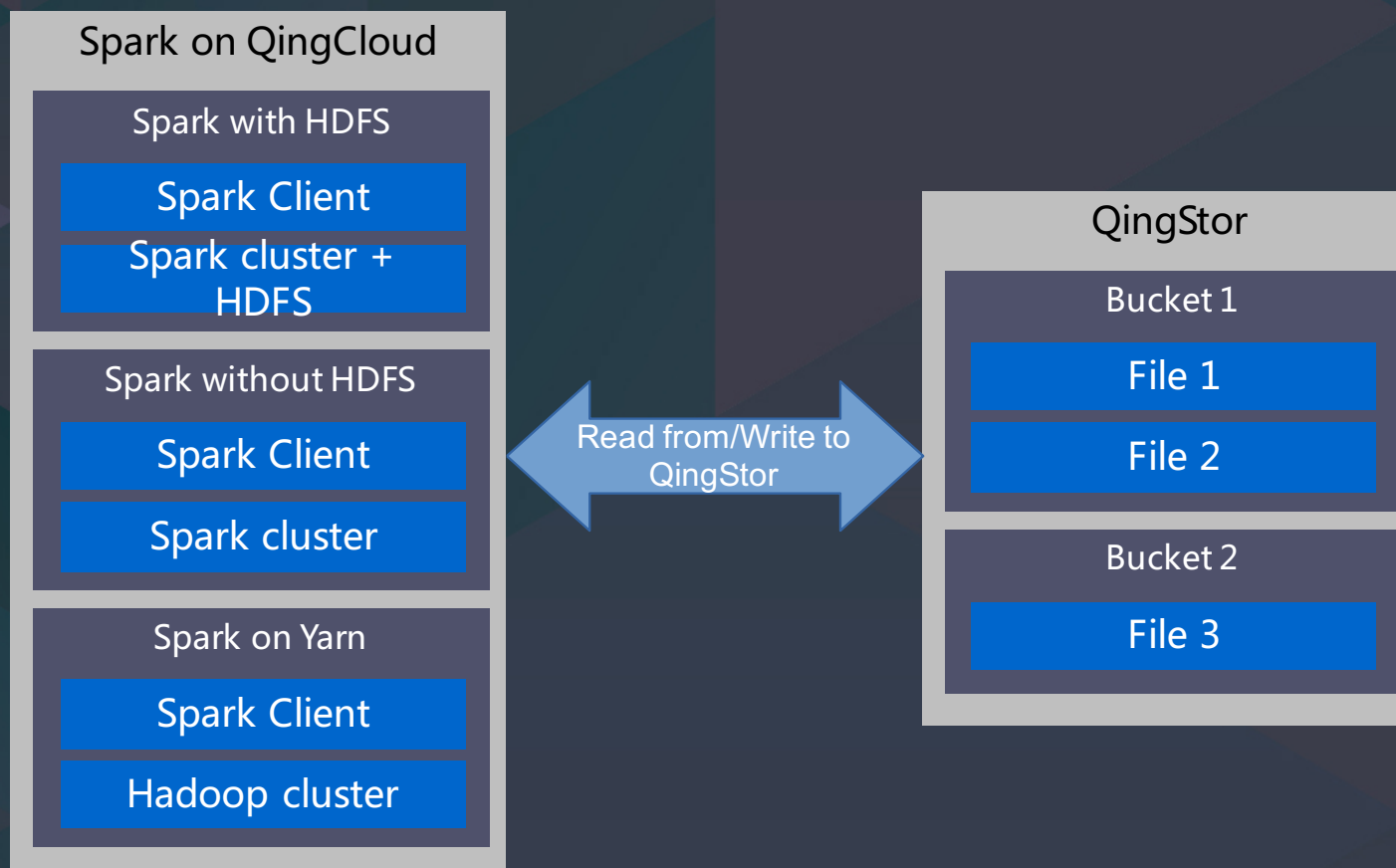
ElasticSearch 与 QingStor 对象存储集成场景



ElasticSearch 与 QingStor 对象存储集成场景

- ▶ 创建/查看/删除 repository
- ▶ 创建/查看/删除 snapshot
- ▶ 恢复snapshot到原集群（可以指定恢复某些或者全部index）
- ▶ 恢复snapshot到一个新的ES集群(可以属于另一个区)

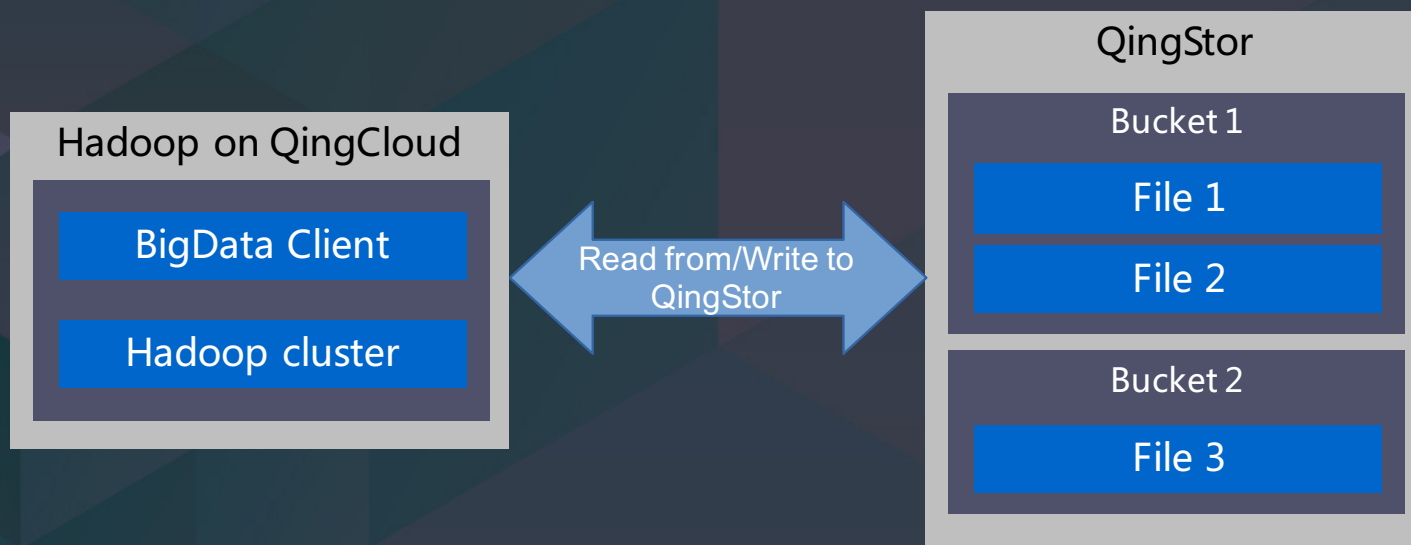
Spark 与QingStor 对象存储集成场景



Spark 与 QingStor™ 对象存储集成场景

- ▶ 在 Spark 中读取 HDFS 上的文件后将其存储到 QingStor™ 对象存储中
- ▶ 在 Spark 中读取 QingStor 对象存储上的文件，处理后再存储到 HDFS
- ▶ 在 Spark 中读取 QingStor 对象存储上的文件，处理后将结果存回 QingStor™ 对象存储
- ▶ 直接将 Spark 的计算结果存储到 QingStor™ 对象存储

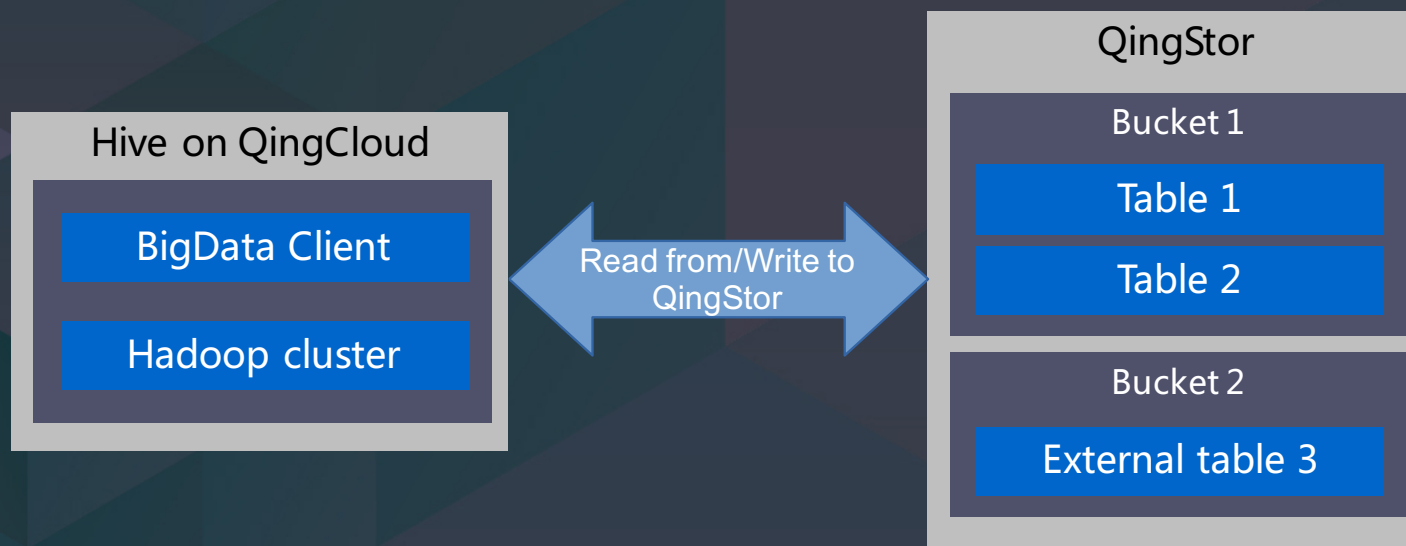
Hadoop 与 QingStor™ 对象存储集成场景



Hadoop 与 QingStor™ 对象存储集成场景

- ▶ 在 QingStor™ 与本地文件系统、HDFS 间读写数据
- ▶ 将 QingStor™ 中的文件作为 mapreduce job 的输入
- ▶ 将 QingStor™ 作为 mapreduce job 的输出

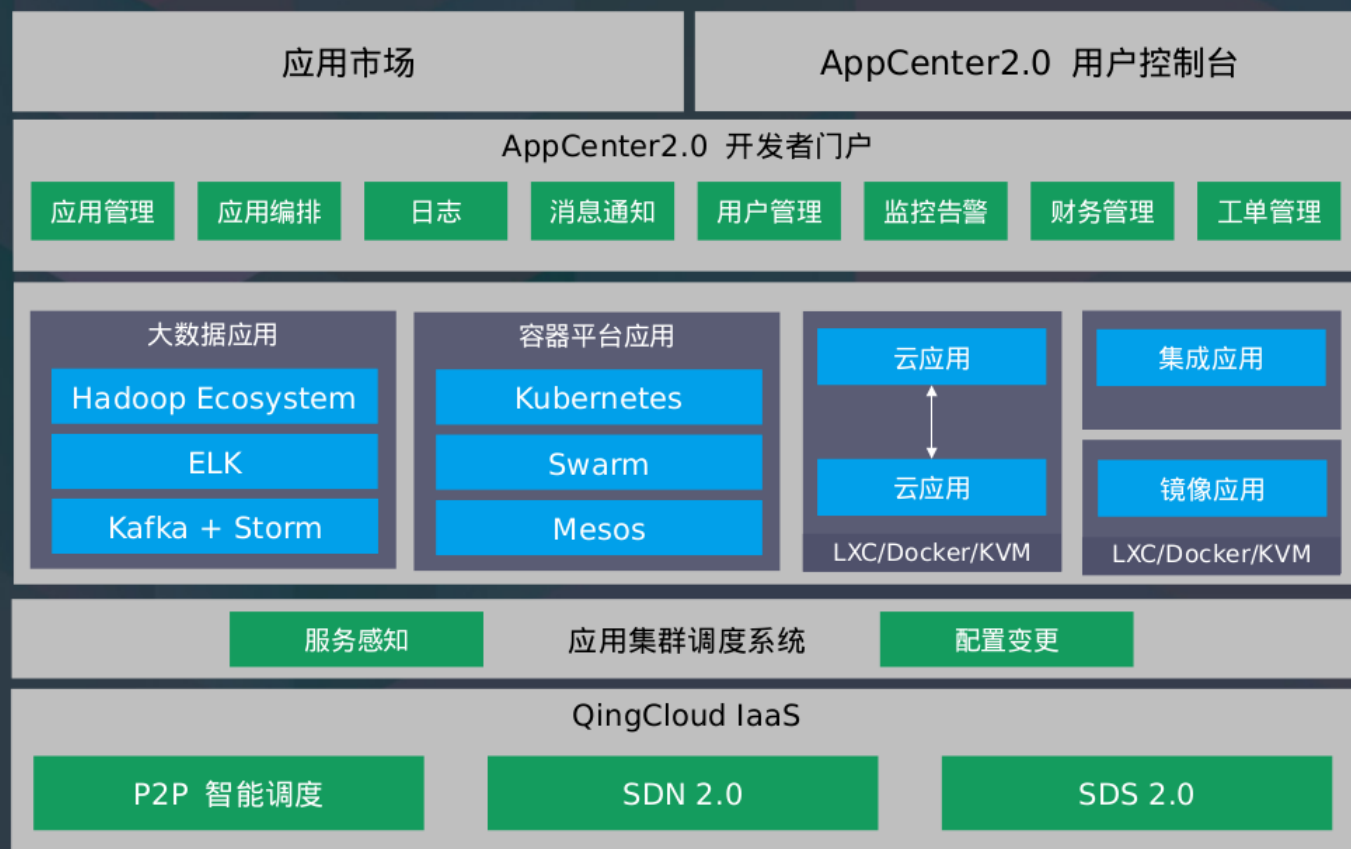
Hive 与 QingStor™ 对象存储集成场景



Hive 与 QingStor™ 对象存储集成场景

- ▶ 将 QingStor™ 对象存储做为 Hive 中 Database 的默认存储引擎，并可创建、查询、操作以 QingStor™ 对象存储作为存储的表
- ▶ 将 HDFS 做为 Hive 中 Database 的默认存储引擎，创建基于 QingStor™ 对象存储的外部表
- ▶ 将 QingStor™ 对象存储中的数据导入以 HDFS 为存储的 Hive 表

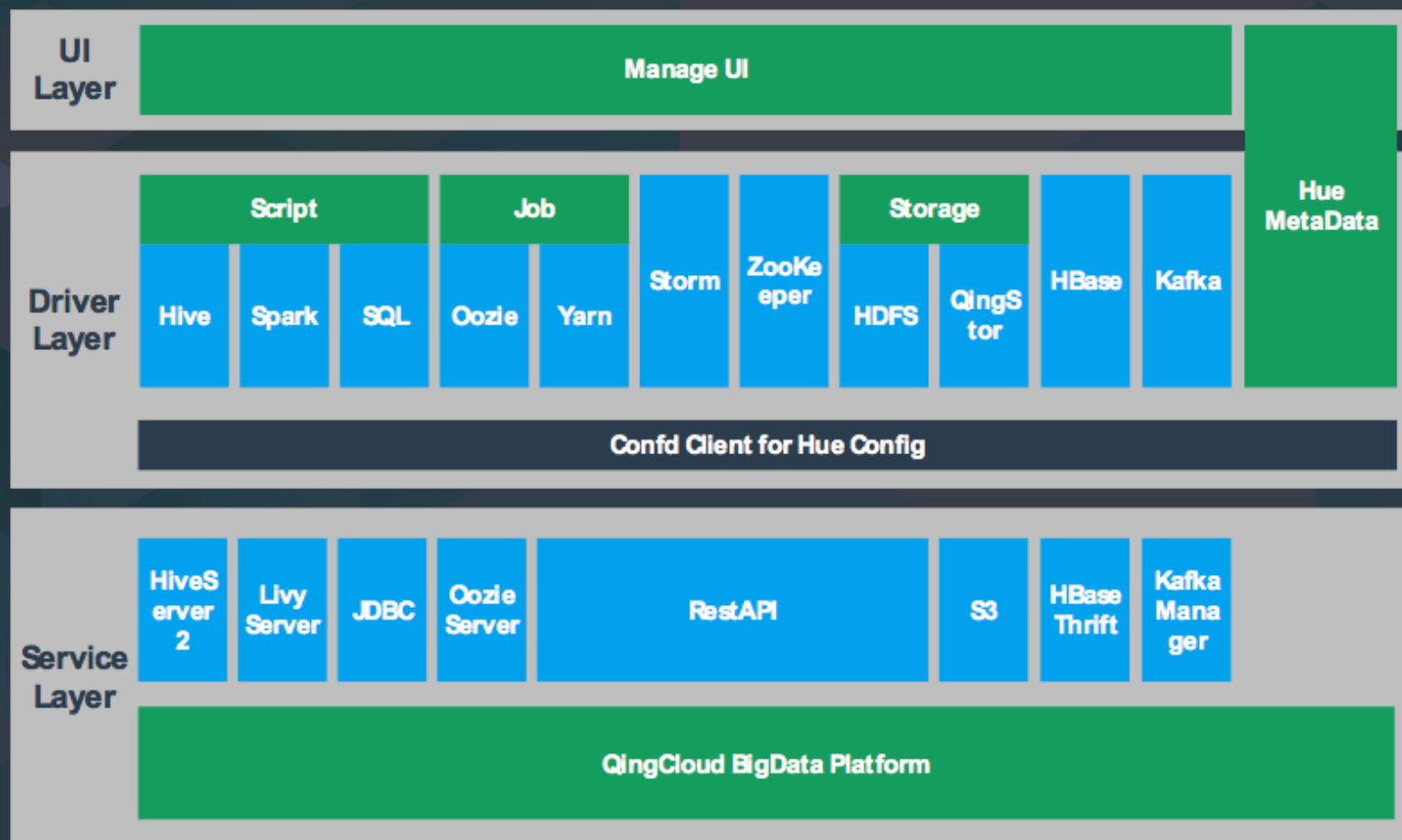
大数据平台 on AppCenter 2.0



What's Next?

HUE
Hadoop 2.7
HBase 1.3
Spark 2.1.1
ELK 5.0
Zeppelin
Kylin
Presto
Impala
Flink
TensorFlow
Caffe
Cassandra
Neo4j
.....

大数据管理平台



大数据平台 on AppCenter 2.0 - 大数据云应用市场

云应用



Mongo on TerarkDB

使用可检索压缩技术，压缩率更高，随机读更快，成本更低

云应用



GBase 8a MPP Cluster 列存...

南大通用面向海量数据分析自主研发的高性能列存数据库

云应用



MySQL on TerarkDB

使用可检索压缩技术，压缩率更高，随机读更快，成本更低

云应用



QingCloud - Apache ZooKeeper...

原生 Apache ZooKeeper 云服务，支持弹性伸缩

云应用



GBase 8a 列存储数据库

南大通用面向海量数据分析自主研发的高性能数据库

大数据平台 on AppCenter 2.0 - 大数据云应用市场

云应用  Yonghong Z-Suite 本应用是一款商务智能软件，帮助用户对数据分析和展现的需求。	 iQuicker 一个全面的企业协同管理平台，敏捷、高效、创新、精细	云应用  小波数据 基于用户行为的数据分析产品，用数据驱动增长	云应用  数据仓库 HashData 高性能 MPP 数据仓库
云应用  Nextcloud on QingCloud A... 后端存储基于 QingStor 对象存储的企业级网盘	 商机无忧 企业线索/商机/销售管理系统	 Testalent 在线笔试平台 Testalent 为企业提供高效的在线笔试方案，简洁的产品交互，流程化的贴心	 掌测 掌测-App测试专家 掌动智能，拥有国家质检总局（国家认监委）颁发的测试资质，无缝匹配App

大数据平台 on AppCenter 2.0 - 大数据云应用示例

集群详情

  新增节点  删除

合计: 7 每页: 10

☐	节点 ID	名称	角色	节点状态	服务状态	配置	私有网络	内网 IP	监控
☐	cln-o2zu6wdr	无	从节点	 活跃	 正常	1核2G 30G	vxnet-f1wa5ox	192.168.0.5	
☐	cln-f9b5lu9r	无	从节点	 活跃	 正常	1核2G 30G	vxnet-f1wa5ox	192.168.0.7	☐
☐	cln-3yporz5y	无	从节点	 活跃	 正常	1核2G 30G	vxnet-f1wa5ox	192.168.0.8	☐
☐	cln-g0mweaz8	无	从节点	 活跃	 正常	1核2G 30G	vxnet-f1wa5ox	192.168.0.6	☐
☐	cln-1u723tml	无	Spark主节点	 活跃	 正常	1核2G 10G	vxnet-f1wa5ox	192.168.0.3	☐
☐	cln-tyebvd8s 	无	Spark client节点	 活跃	无监控	1核2G 10G	vxnet-f1wa5ox	192.168.0.2	☐
☐	cln-mj5raqfm	无	HDFS主节点	 活跃	 正常	1核2G 10G	vxnet-f1wa5ox	192.168.0.4	☐

大数据平台 on AppCenter 2.0 - 大数据云应用示例

节点

配置参数

配置参数是在部署应用时填写，成为集群配置项的变量。有的配置项是公共的，有的作用于其中的一个或多个角色。您可以在这里修改参数，以更新集群配置。

公共参数

参数 [修改属性]	值	描述	范围
<code>spark.worker.cleanup.enabled</code>	true	定期清理应用work目录，运行中的application不会被清理。	
<code>spark.worker.cleanup.interval</code>	28800	清理应用work目录的时间间隔，以秒为单位，默认为28800秒(8小时)	
<code>spark.worker.cleanup.appDataTtl</code>	86400	保留worker上应用work目录的时间，以秒为单位，默认为86400秒(24 小时)	
<code>hadoop.proxyuser</code>		Hadoop代理用户	
<code>hadoop.proxyuser.[hadoop_proxyuser].hosts</code>	*	Hadoop代理用户能代理哪些hosts	
<code>hadoop.proxyuser.[hadoop_proxyuser].groups</code>	*	Hadoop代理用户能代理指定host中的哪些groups	
<code>dfs.datanode.handler.count</code>	10	Data node 节点服务线程数	

大数据平台 on AppCenter 2.0 - 大数据云应用示例

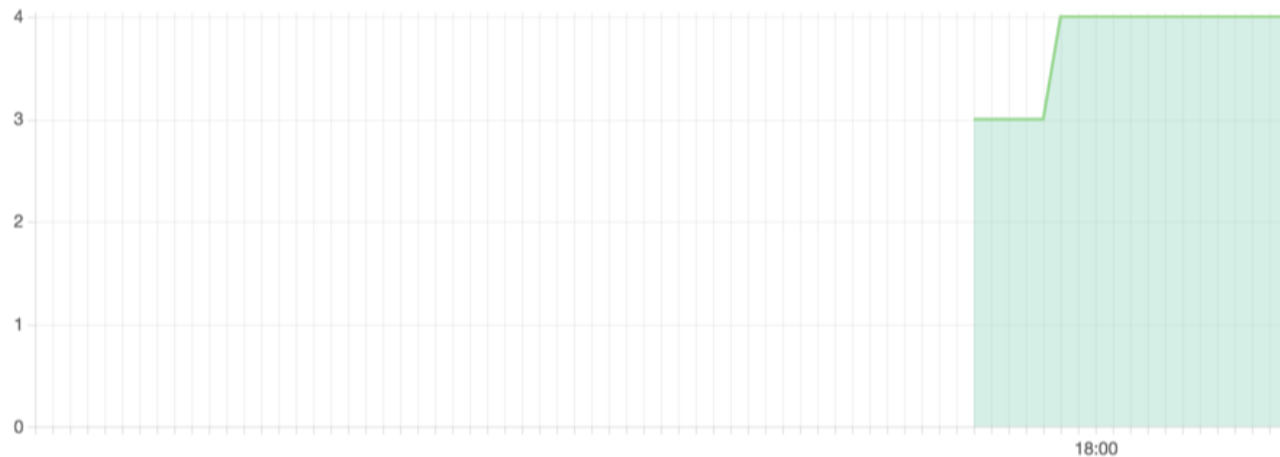
Spark 主节点监控

WORKERS

单位: 间隔: 5m

监控项: ● Total ● Alive

关



大数据平台 on AppCenter 2.0 - 大数据云应用示例

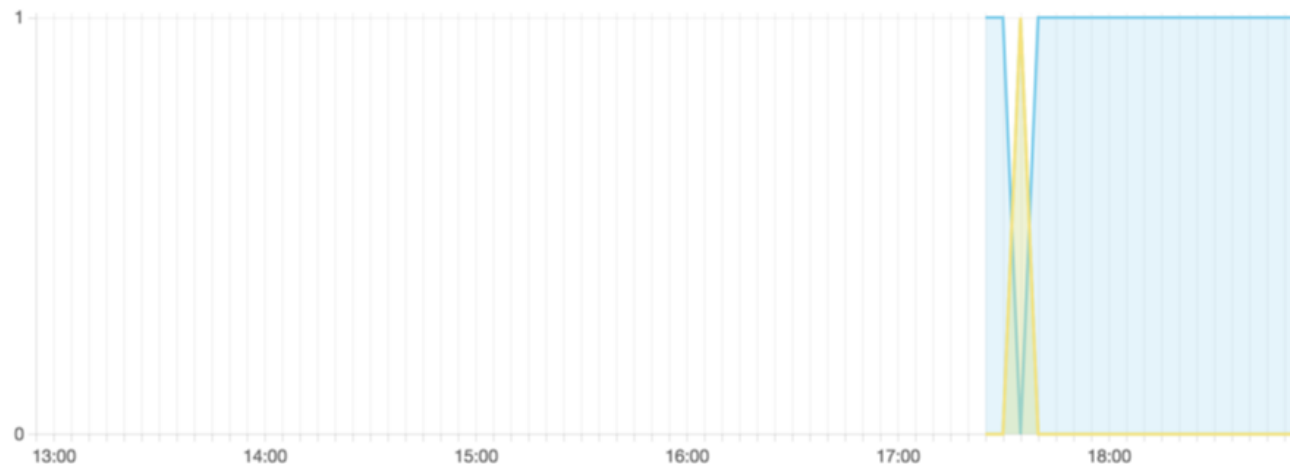
Spark 从节点监控

计算

单位: 间隔: 5m

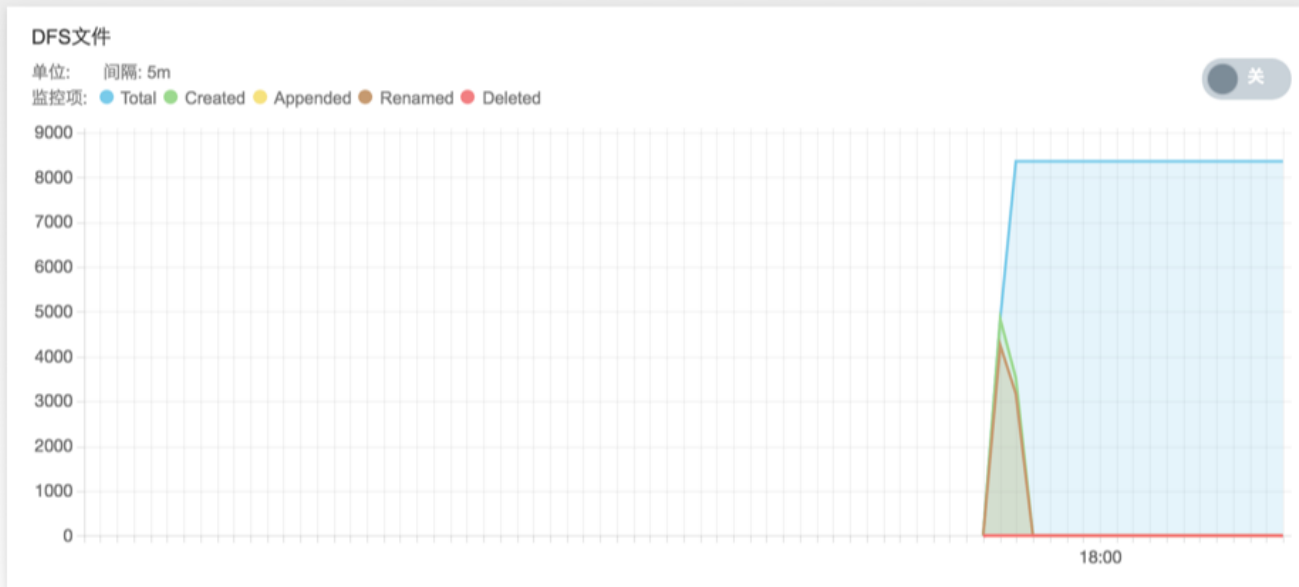
监控项: CoresFree CoresUsed Executors

关



大数据平台 on AppCenter 2.0 - 大数据云应用示例

HDFS 主节点监控



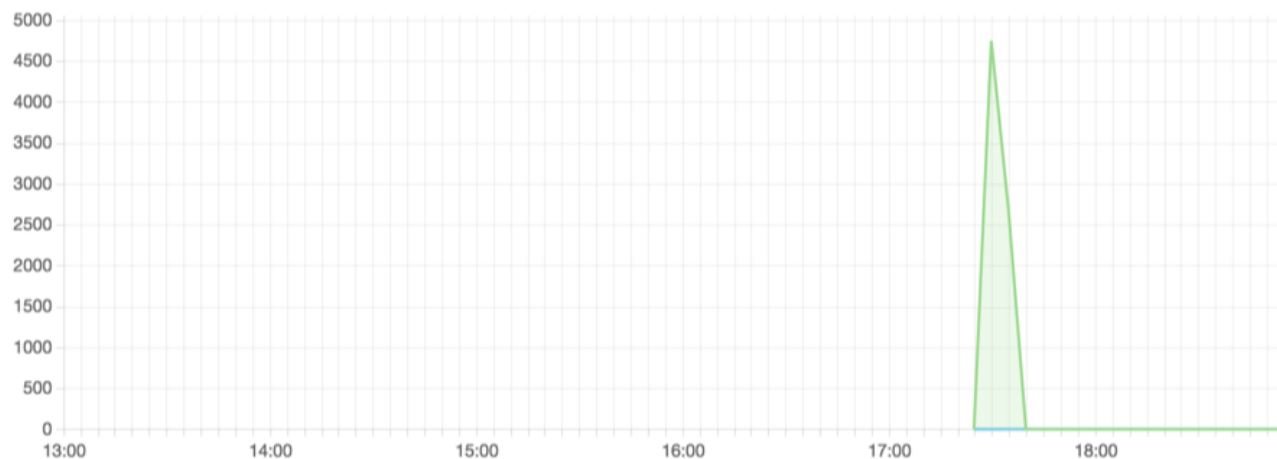
大数据平台 on AppCenter 2.0 - 大数据云应用示例

HDFS 从节点监控

DFS块

单位: 间隔: 5m

监控项: Read Written



总结

- ▶ 千差万别的业务场景、层出不穷的大数据技术，如何找到合适的大数据解决方案？
- ▶ QingStor对象存储是云端大数据技术的基石
- ▶ AppCenter提供了快速便捷搭建云端大数据解决方案的：
 - ▶ 可能性
 - ▶ 多种选择
 - ▶ Building block



实践课堂 / 第3季 / 上海站

QingCloud Workshop Season 3 / Shanghai

第6期 / 2017.6.10