小象科技
ChinaHadoop.cn

Spark 其他开源系统

董西成
2017年5月

主要内容

1. Alluxio
2. Spark JobServer
3. Zeppelin
4. 总结

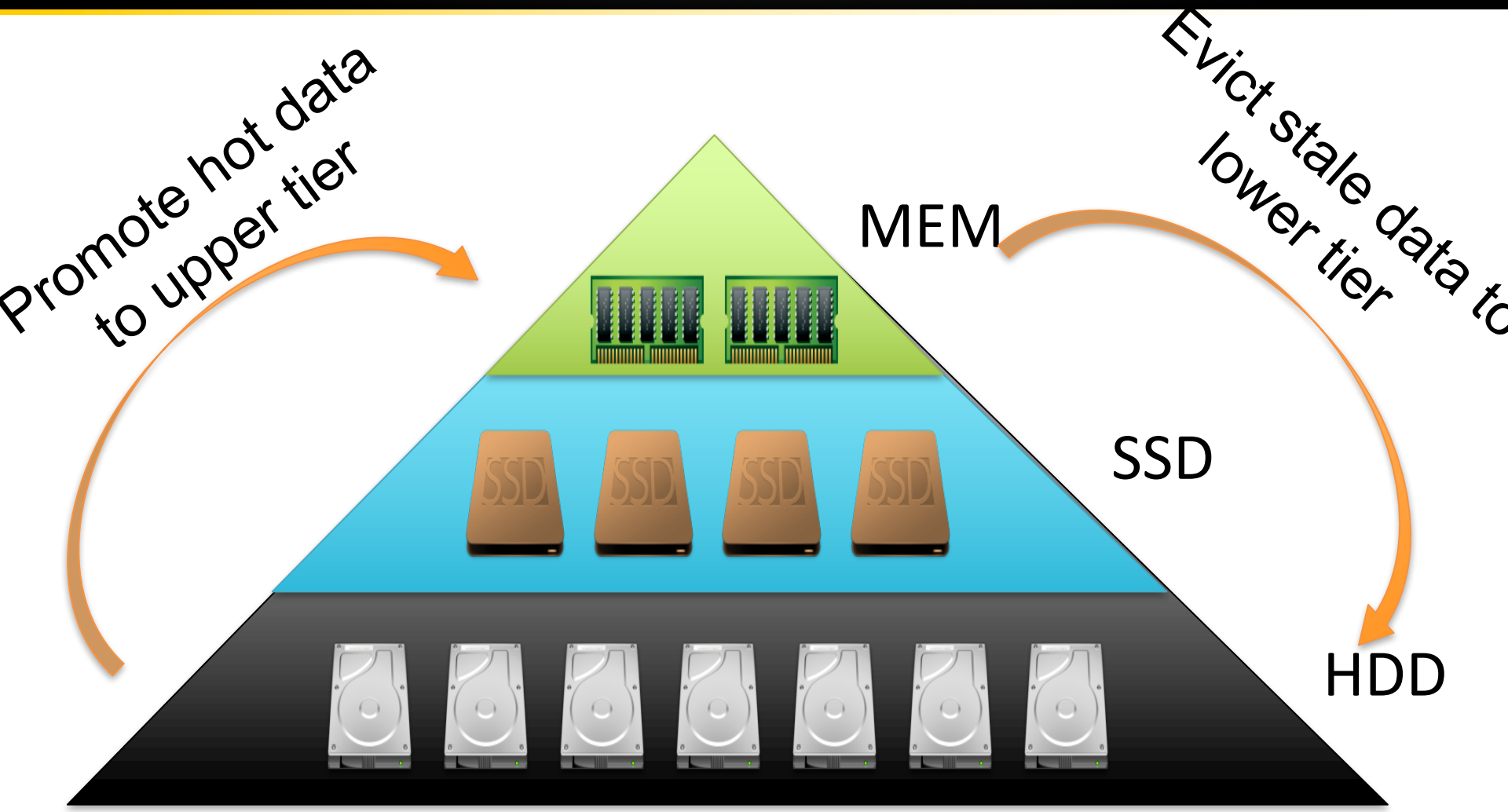
主要内容

1. Alluxio
2. Spark JobServer
3. Zeppelin
4. 总结

Alluxio是什么？

- 一个虚拟分布式文件系统；
- 架构在底层分布式文件系统和上层分布式计算框架之间的一个中间件；
- 以文件形式在内存或其它存储设施中提供数据的存取服务。
- 官网链接：http://www.alluxio.org/

Tied Storage



Alluxio生态系统

Case Study: Baidu

Baidu Queries Data 30 Times Faster with Alluxio

- Application: Spark
- Alluxio Storage: MEM + HDD
- Backend Storage: Baidu's File System
- 200+ nodes deployment, 2PB+ managed space
- Result: Speeding up data querying by 30x

主要内容

1. Alluxio
2. **Spark JobServer**
3. Zeppelin
4. 总结

什么是Spark JobServer

- “Spark as Service”：针对 job 和 contexts 的各个方面提供了 REST 风格的 api 接口进行管理（用REST提交、杀死以及管理spark作业）
- 通过长期运行的job contexts支持亚秒级别低延迟的任务
- 可以通过结束 context 来停止运行的作业(job)
- 分割 jar 上传步骤以提高 job 的启动速度
- 异步和同步的 job API，其中同步 API 对低延时作业非常有效
- 支持 Standalone Spark 和 Mesos、yarn
- 对RDD或DataFrame对象命名并缓存，通过该名称获取RDD或DataFrame。这样可以提高对象在作业间的共享和重用
- 官网链接：https://github.com/spark-jobserver/spark-jobserver

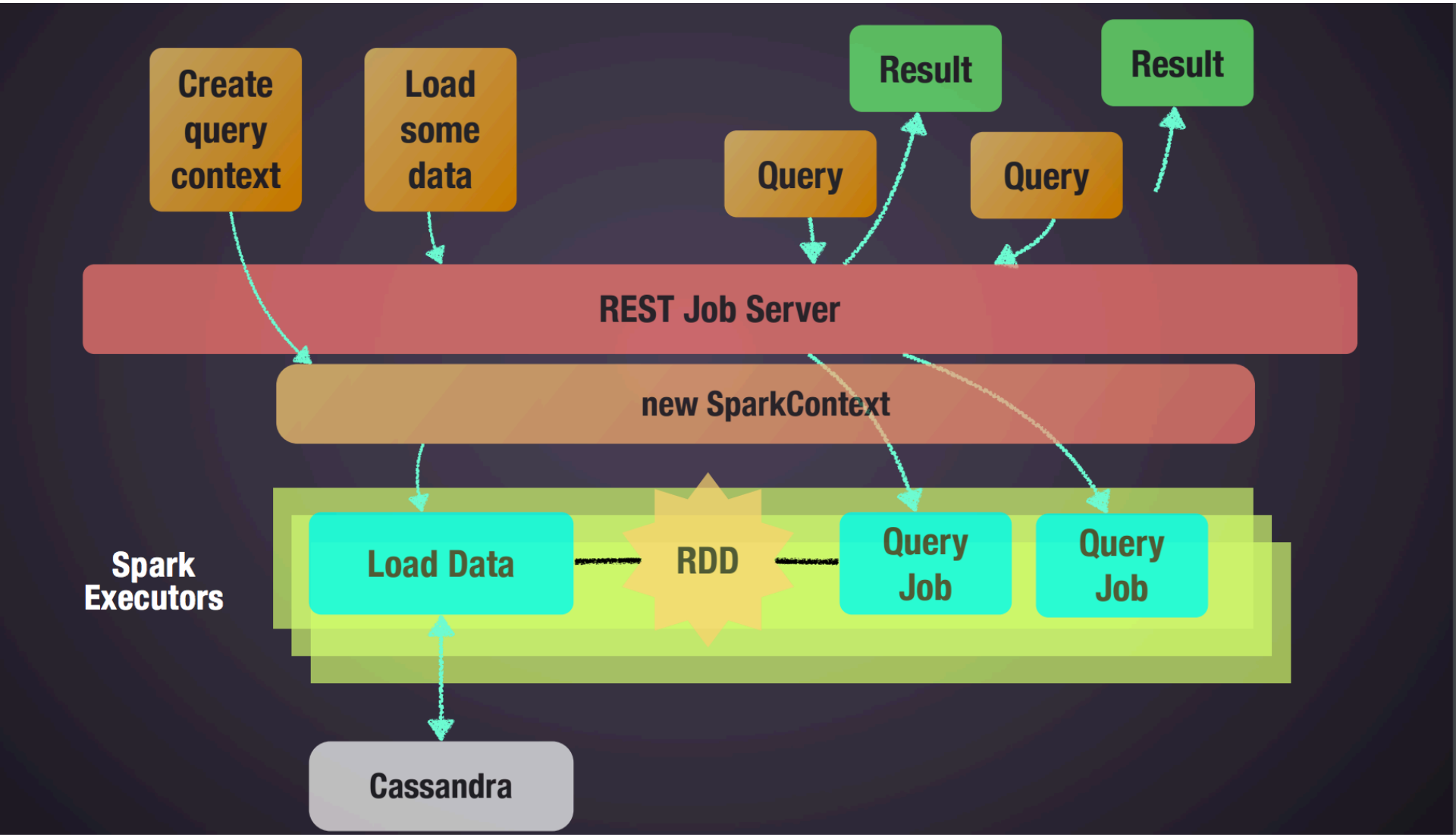
Spark JobServer Job实例

```
/**
 * A super-simple Spark job example that implements the SparkJob trait and
 * can be submitted to the job server.
 */
object WordCountExample extends SparkJob {
  override def validate(sc: SparkContext, config: Config): SparkJobValidation = {
    Try(config.getString("input.string"))
      .map(x => SparkJobValid)
      .getOrElse(SparkJobInvalid("No input.string"))
  }

  override def runJob(sc: SparkContext, config: Config): Any = {
    val dd = sc.parallelize(config.getString("input.string").split(" ").toSeq)
    dd.map((_, 1)).reduceByKey(_ + _).collect().toMap
  }
}
```

Spark JobServer Job提交

```
✦  curl --data-binary @../target/mydemo.jar localhost:8090/jars/demo
OK[11:32 PM] ~

✦  curl -d "input.string = A lazy dog jumped mean dog" 'localhost:8090/jobs?
appName=demo&classPath=WordCountExample&sync=true'
{
  "status": "OK",
  "RESULT": {
    "lazy": 1,
    "jumped": 1,
    "A": 1,
    "mean": 1,
    "dog": 2
  }
}
```

Spark JobServer加速Job运行

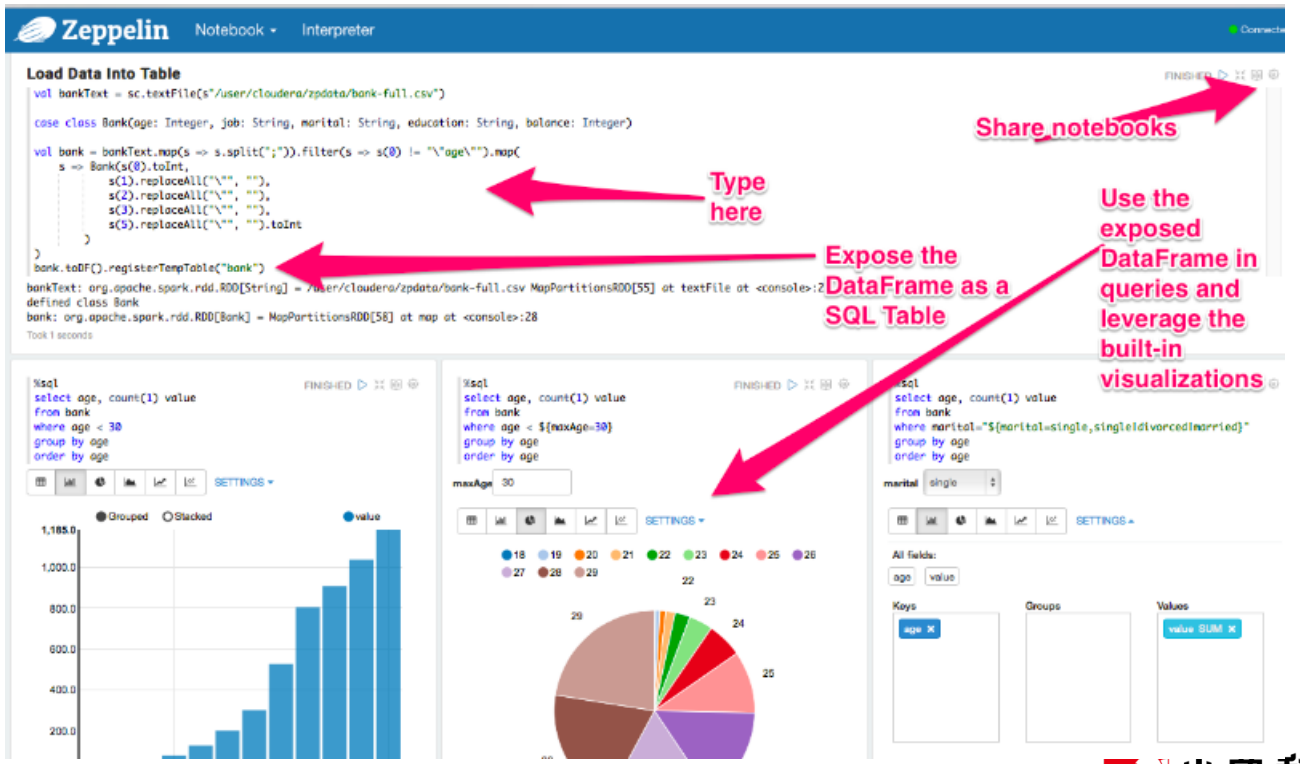


主要内容

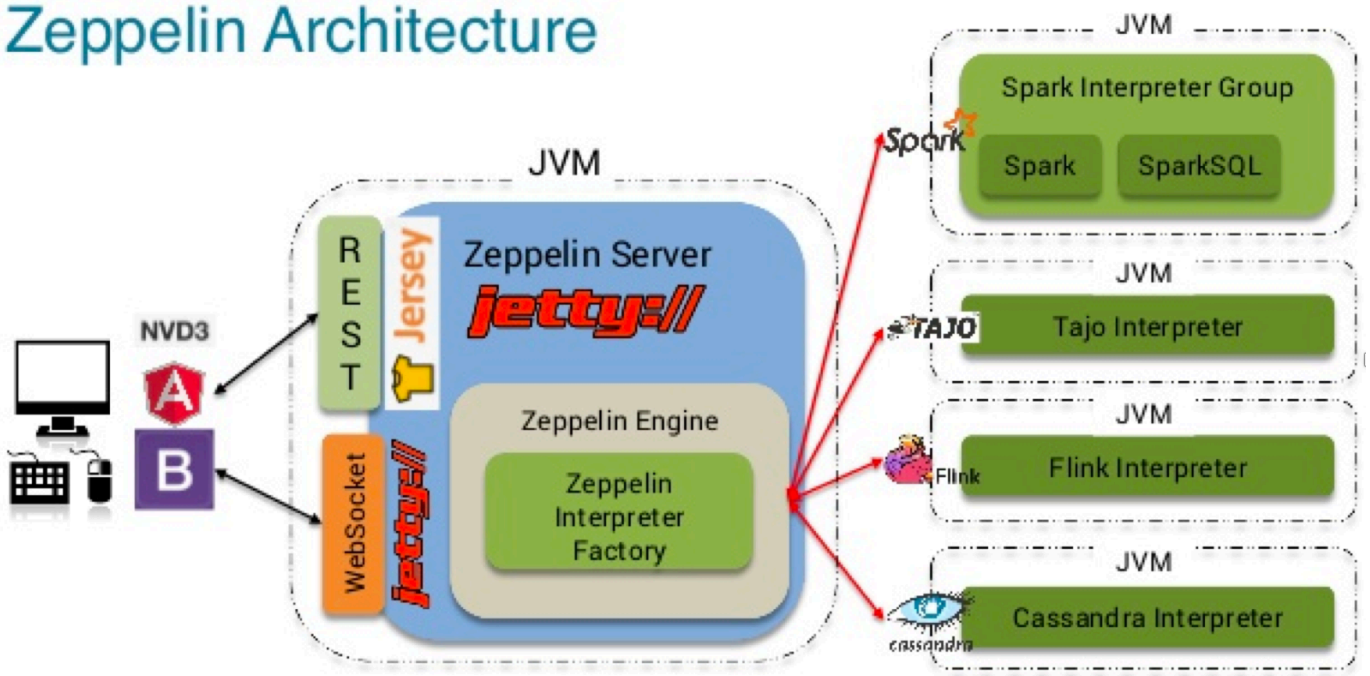
1. Alluxio
2. Spark JobServer
3. **Zeppelin**
4. 总结

Zeppelin: 基于web的一站式数据科学平台

- 数据探索和发现
- 官网链接： http://zeppelin.apache.org/
- 与Spark和Hadoop深度集成
- 插件化解释器

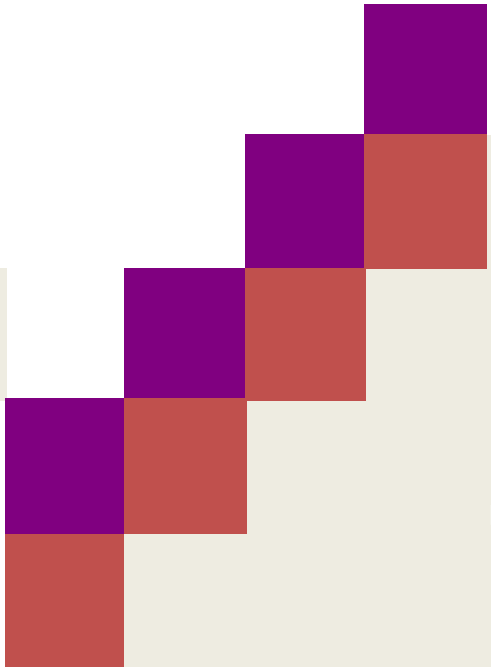


Zeppelin: 插件化架构



hadoop123：董西成的微信公众号

专注于Hadoop/spark等大数据相关技术的分享



Q&A?

Thank You !