爱奇艺

悦 享 品 质

爱奇艺Spark平台构建之路

2016-11-25

Who Am I?

- 顾亮亮
- 上海交大：2013毕业
- SAP：机器学习算法开发 1年
- 爱奇艺：大数据平台开发 3年
- 关注：分布式计算框架，流式计算

Agenda

- 第一阶段：裸用Spark
- 第二阶段：平台化管理Spark
- 第三阶段：不用写Spark
- 第四阶段：不只是Spark
- 平台构建经验

第一阶段：裸用Spark (2014)

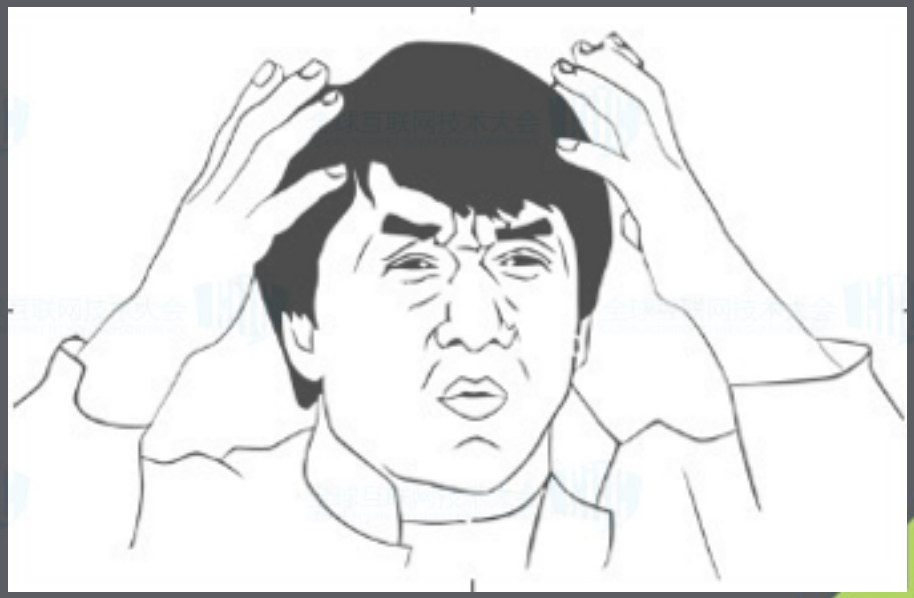
- Spark-1.2.1、Spark-1.3.0
- Standalone x 4 + Spark on YARN x 1
- 入口机
- 100个Batch任务/天 + 0个Streaming任务

增值服务

- 提供优化过的Spark默认参数
  - spark.yarn.executor.memoryOverhead=2048
  - …
- 提供库：解决用户问题
  - HBase Kerberos问题
  - …
- 提供各种文档
  - 例子
  - 使用手册
  - 调优手册
  - Debug手册
  - …

问题

- Standalone扩展不方便
- 入口机缺少HA
- 测试升级维护比较麻烦
- 缺少监控、重试、调试、查错、调度、管理等机制
- 运维占用大量时间
  - 为什么我的Spark任务失败了?
  - 为什么任务卡住了?
  - 为什么这么多failed task?
  - 为什么无法提交任务?
  - 为什么程序抛了这个错?
  - …

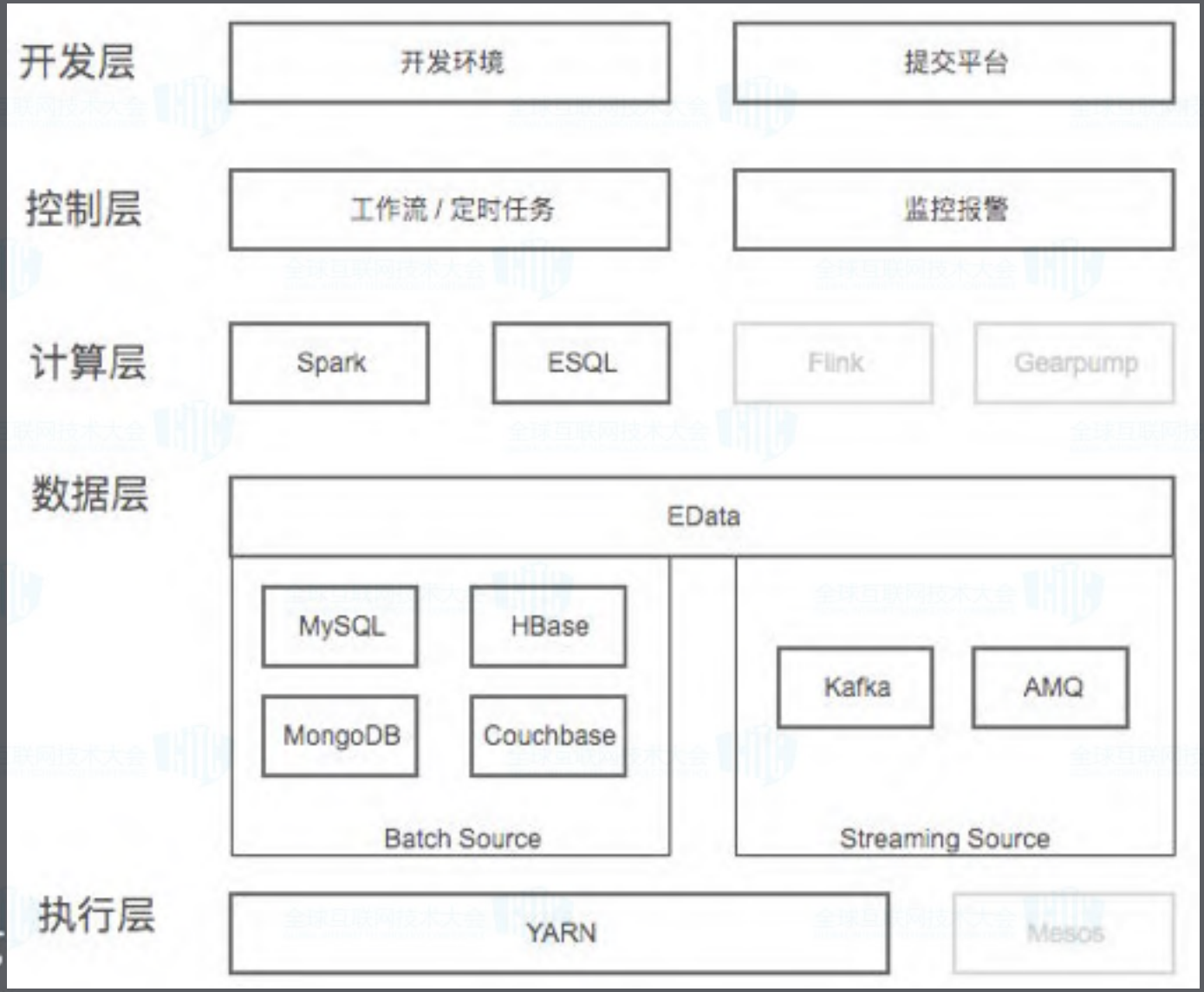


第二阶段：平台化管理Spark (2015 - 2016)

- Spark-1.4.1、Spark-1.5.2、Spark-1.6.2、Spark-2.0.1
- 使用平台提交到Yarn上运行，Yarn集群 x 4
- 2015
  - 500个Batch任务/天 + 100个Streaming任务
- 2016
  - 2K个Batch任务/天 + 500个Streaming任务
  - 15K个核 + 50T内存


汉密尔顿2014年方向盘

平台架构



增值服务

提交运行Spark程序

网页提交

命令行提交 (Linux & Windows)

```
Error, please input an europa command.
Usage: europa <command> [options]
Commands:
       help: output the help information for europa command
       init: please execute this command firstly, initial local europa environment for user
        app: query, manage(create, run...) Spark app images
      batch: manage spark batch images
  streaming: manage spark streaming task
        job: europa job(s)/task batch and streaming) lifecycle manage command, like kill and query job etc
    project: manage europa project
      pipe: manage(create, run...) pipeline tasks
    version: get the local europa client version
Use 'europa help' for more information.
```

Maven插件提交

Java SDK提交

监控报警 & 自动重启

- Streaming任务自动重启
- 自动错误诊断（基于规则）
  - Pending Batch超过1h
  - Kafka没有流量超过0.5h
  - Spark Stage卡住超过10min
- 自定义报警
  - Kafka QPS超过20w十分钟（需要申请更多计算资源）
- 资源使用异常检测（基于规则）
  - 申请1000个core，但是最大的并行度只有10

统一API访问不同数据库

- 问题：访问数据源的库
  - 官方提供：MySQL， Hive， Kafka
  - 第三方提供：Mongodb，Couchbase, HBase
  - 没有很好的开源实现：ActiveMQ
  - API不一致，功能也不一样，有些库还有Bug
- 包装成统一的API并发布

其他增值服务

- 高可用
- 定时任务
- 工作流
- 进一步优化SparkConf
  - 1个核：-XX:+UseSerialGC
  - n个核：-XX:+UseParallelGC -XX:ParallelGCThreads=#n
- 团队协作
- 版本管理
- 统计
- 计算资源的管理
- 下载YARN日志并保留7天
- 收集分析程序Metrics （JVM）
- 分布式Profiling（Flame Graph)

问题

- Spark学习成本很大
  - Scala、Java、Python、R
  - RDD、DataFrame、DataSet、DStream
  - DAG、Driver、Executor、YARN
  - SparkConf各种参数设置
- 需要1-2周才能写好Spark Batch程序
  - 逻辑正确
  - 稳定运行
  - 应对数据量的变化
- 需要近1个月才能写好Spark Streaming程序
  - 能处理流量波动和突增
  - Exactly once语义
  - 支持重启
  - 稳定运行
  - 程序升级

Sebastian Vettel shows his first FIA Super License in 2006.

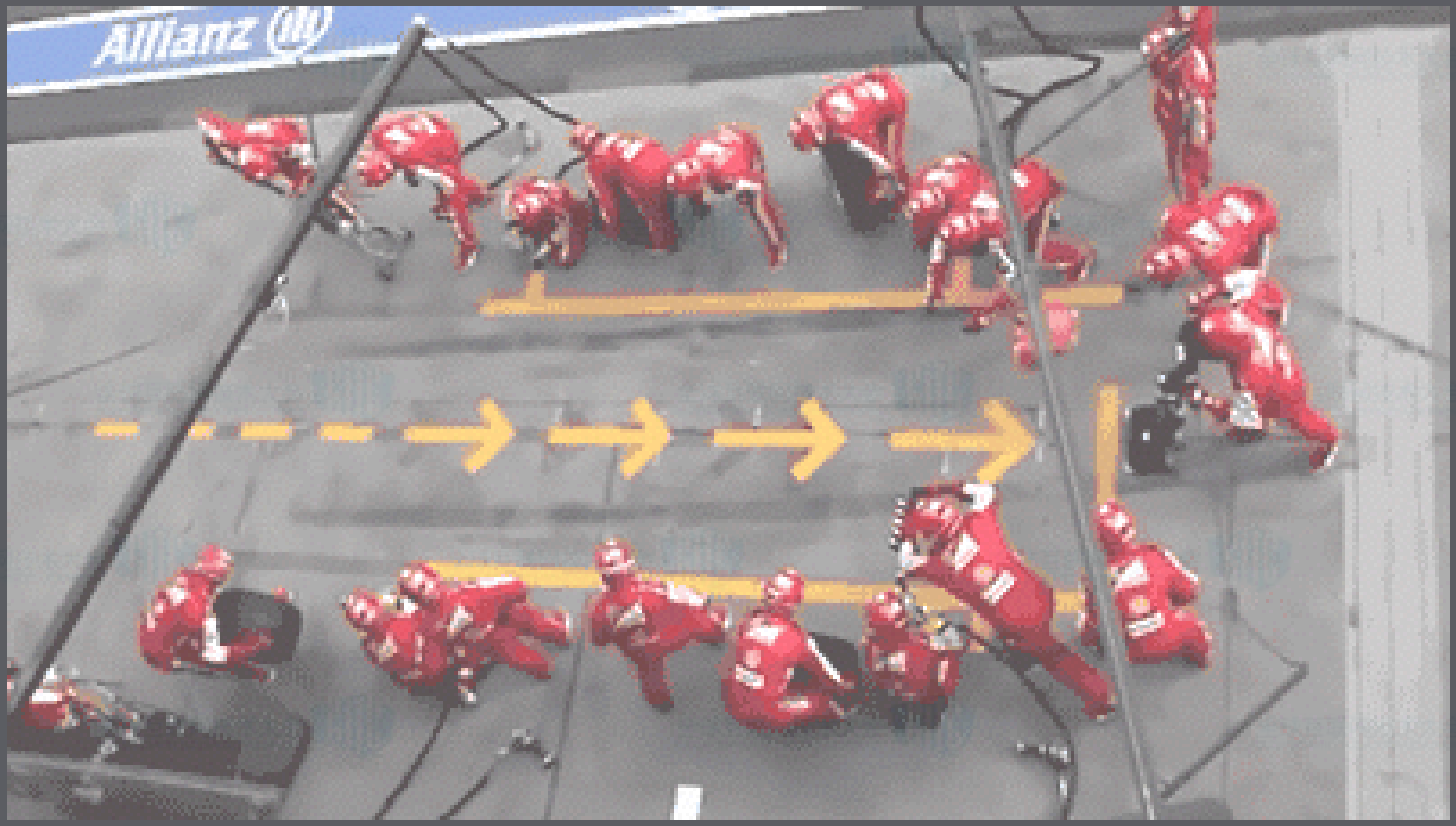
第三阶段：不用写Spark (2016)

- Spark开发工具
  - 用户配置 + SparkSQL
  - 图形界面 + SparkSQL

Example

```
{
  "test": {
    "desc": "从Kafka读取json格式数据，通过正则表达式抓取字符串，并以parquet格式保存到HDFS",
    "strategy": "SparkStreamingStrategy",
    "compositor": [
      {
        "name": "KafkaStreamingCompositor",
        "params": {"metadata.broker.list": "xxx.xxx.xxx.xxx:9092,...", "topics": "vis-nginx-cache-access"}
      },
      {
        "name": "JSONTableCompositor",
        "params": {"tableName": "raw_table", "schema": "dc:string,server:string,raw:string"}
      },
      {
        "name": "SQLCompositor",
        "params": {"sql": "select dc, server, regexp_extract(raw, '((?<= (tvid=)).+?(?= (&|,|\s)))') as tvid from raw_table"}
      },
      {
        "name": "SQLParquetOutputCompositor",
        "params": {"path": "/data/.../vis-nginx-cache-access.parquet/date=${yyyyMMdd}/hour=${HH}", "mode": "Append"}
      }
    ]
  }
}
```

增值服务

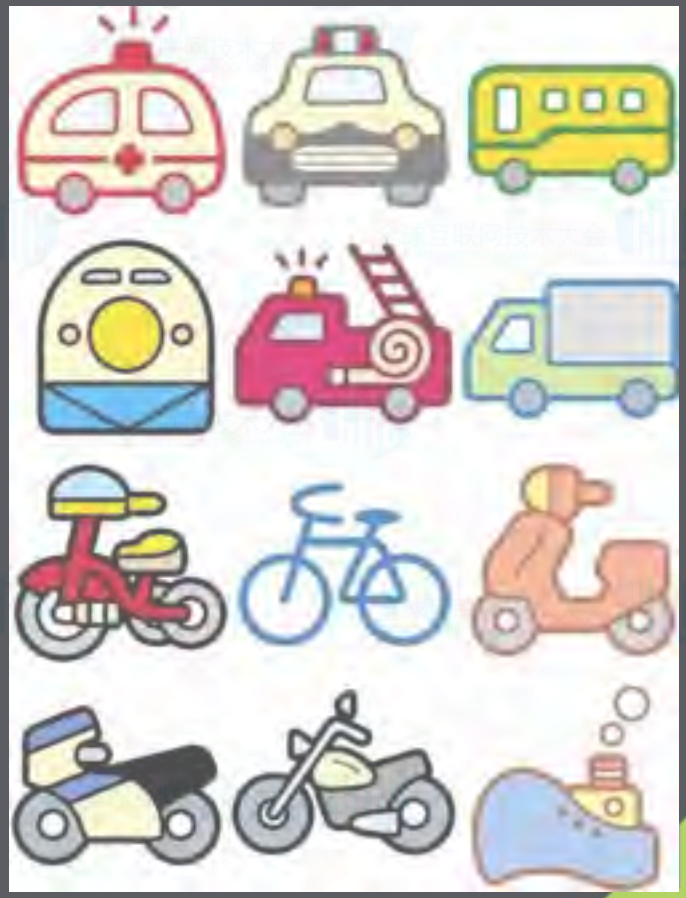


增值服务

- Kafka Offset管理
- 引导用户使用Spark的方式
  - SparkSQL
  - UDF & UDAF
  - 自定义Parser
- 图形界面拖拽 + SQL编辑
- 用户数据源管理

第四阶段：不只是Spark (2017)

- 问题
  - 基于日志真实时间
  - Streaming实时性
  - StreamingSQL
  - …
- 不只是Spark
  - Flink
  - Gearpump
  - Apache Beam
  - …



增值服务

- 一个系统支持运行多个框架
- 一份代码运行到多个框架
- …

平台构建经验

平台提供的增值服务
平台承担的责任

用户自由度
用户编程接口
用户关心的事情

抽象用户共同逻辑，由平台提供
用户只需要写业务相关代码（最好不用写代码），不用关心分布式问题


iQIYI
爱奇艺
悦 享 品 质