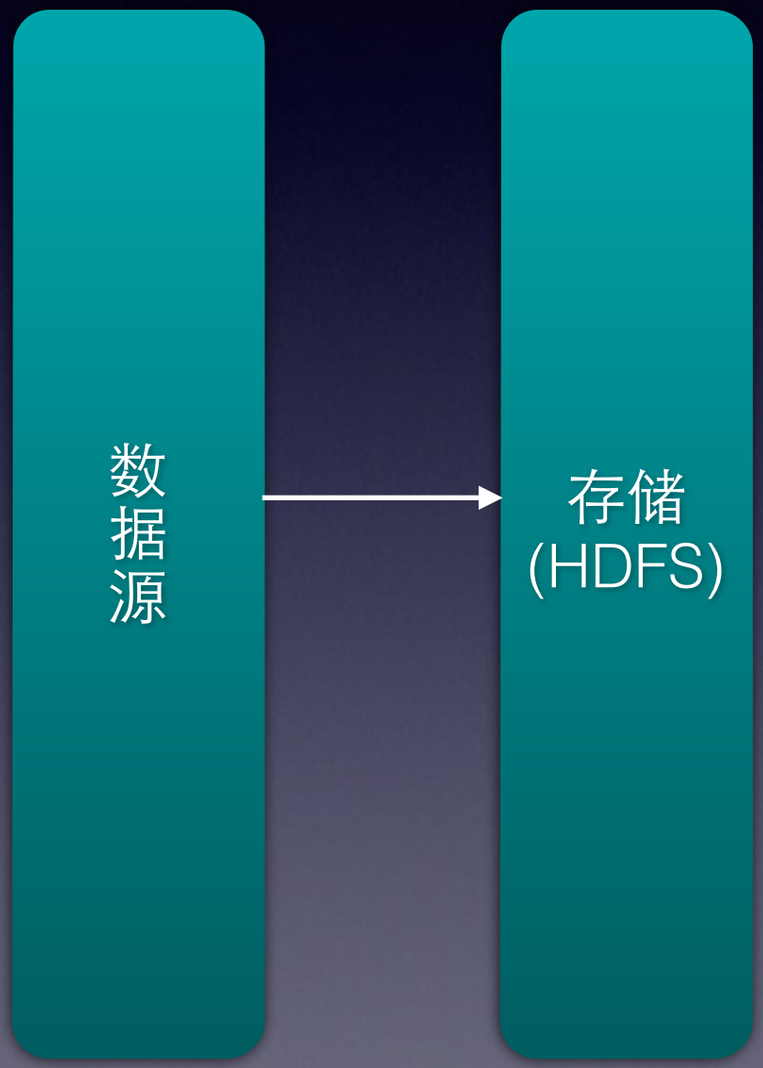
Spark项目架构与实战

陈 超
@CrazyJvm

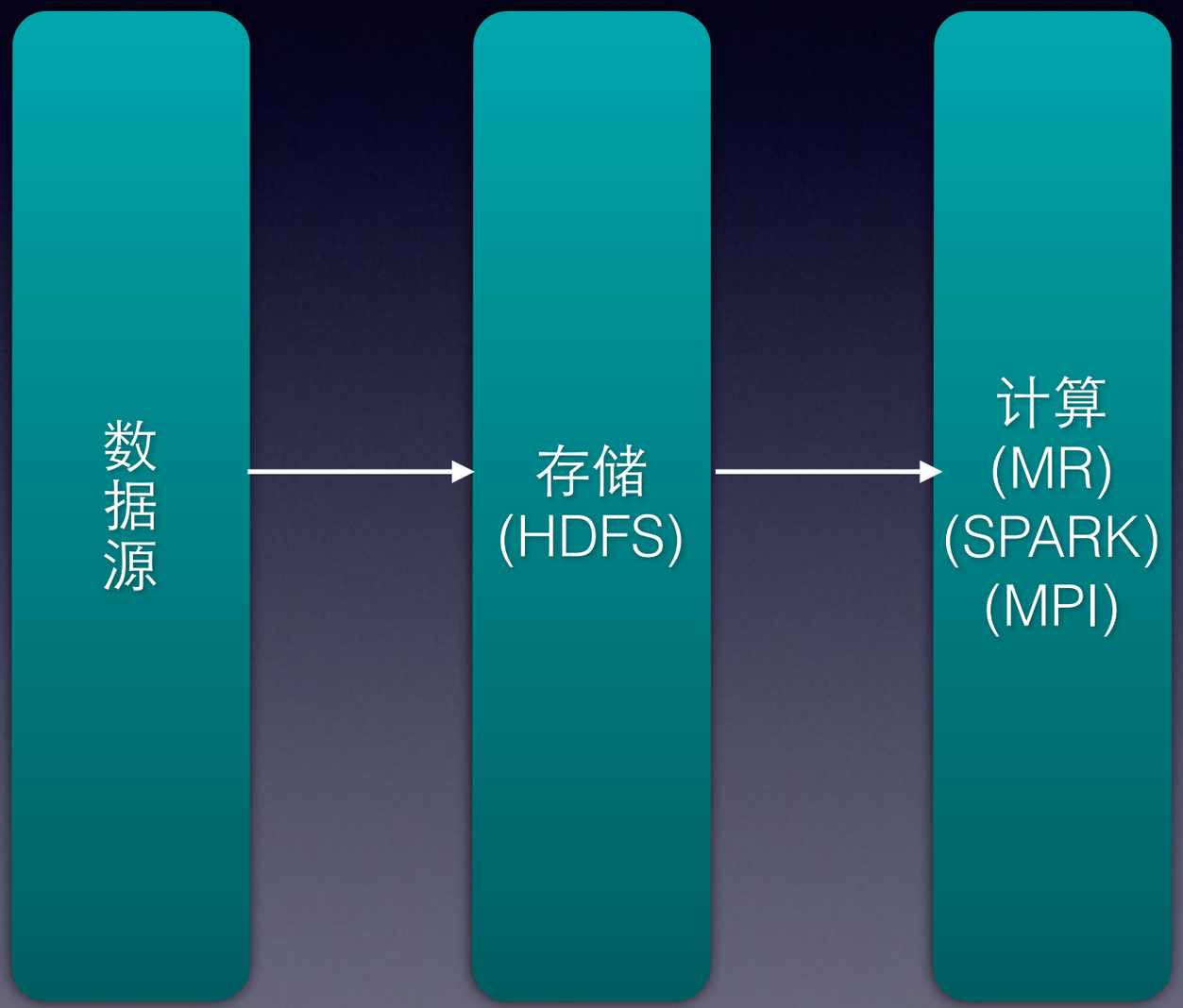
整体流程

数据源

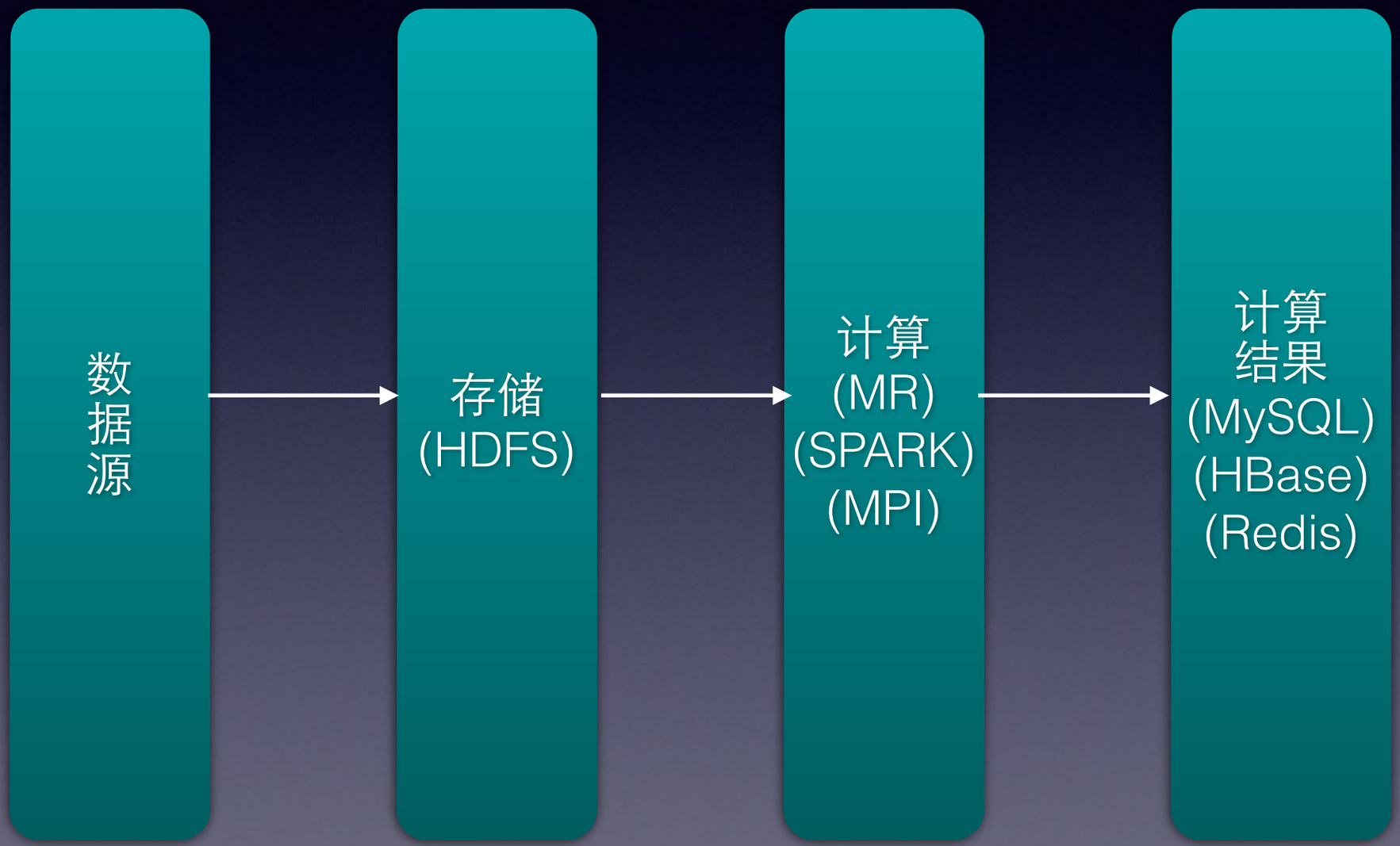
整体流程



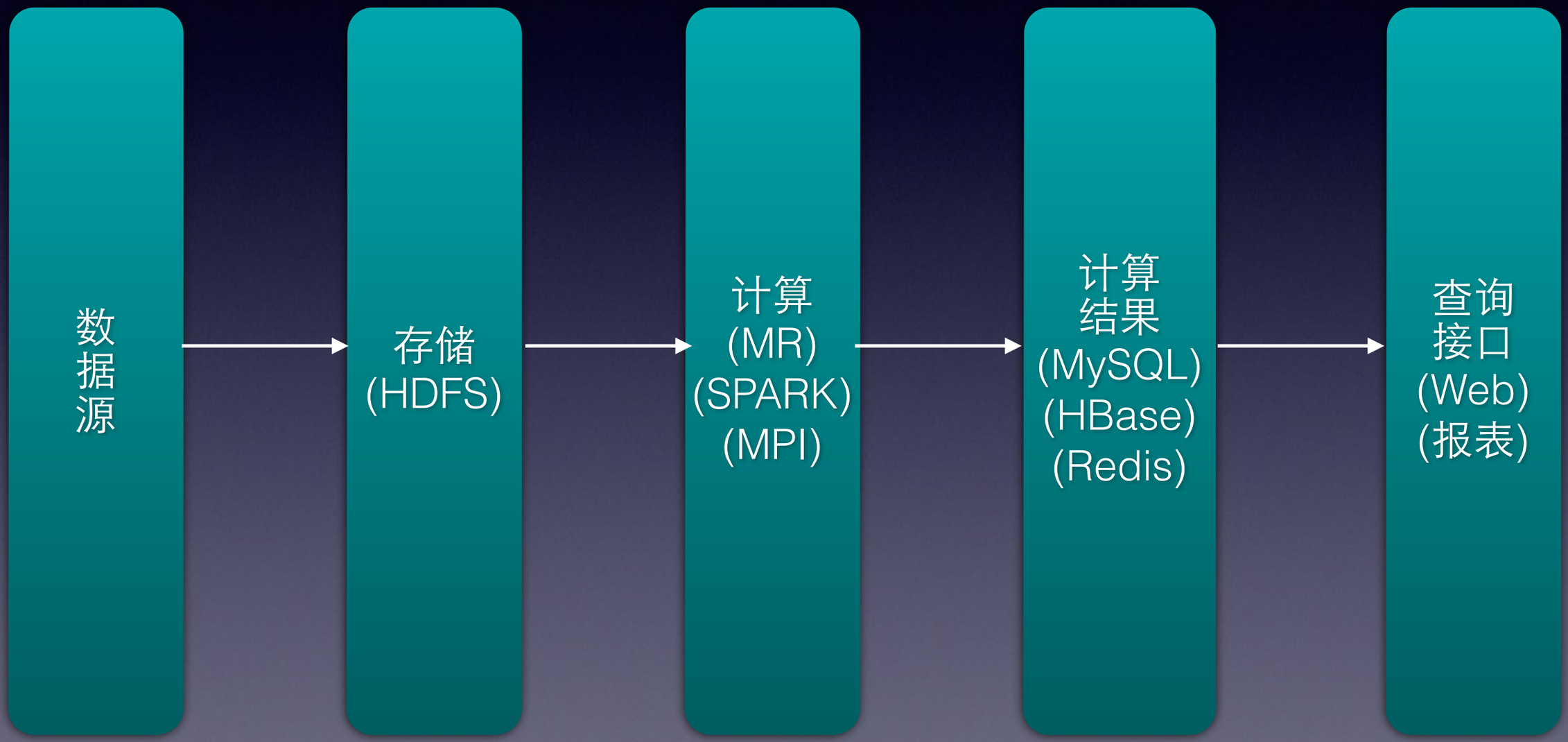
整体流程

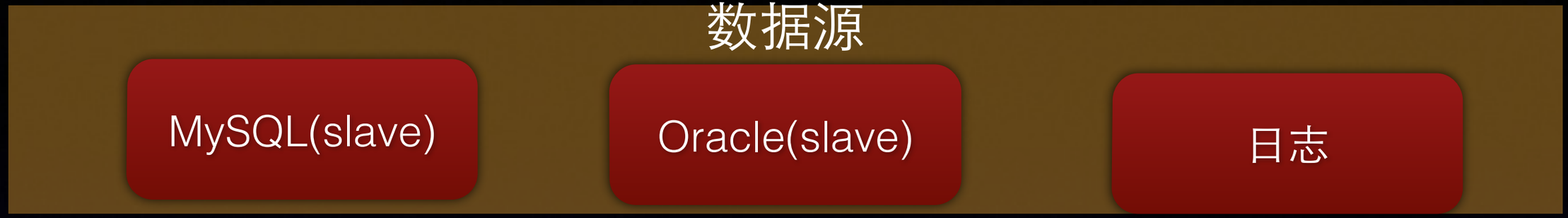


整体流程



整体流程




数据源
MySQL(slave)
Oracle(slave)
日志


数据源
MySQL(slave)
Oracle(slave)
日志
数据同步层
实时同步
非实时同步

数据源

MySQL(slave)

Oracle(slave)

日志

数据同步层

实时同步

非实时同步

HDFS

MR Job

Hive Job

Spark Job

Shark Job

流计算

数据源

MySQL(slave)

Oracle(slave)

日志

数据同步层

实时同步

非实时同步

HDFS

MR Job

Hive Job

Spark Job

Shark Job

流计算

计算结果

MySQL

Oracle

HBase

Redis

GemFire

数据源

MySQL(slave)

Oracle(slave)

日志

数据同步层

实时同步

非实时同步

HDFS

MR Job

Hive Job

Spark Job

Shark Job

流计算

结算结果

MySQL

Oracle

HBase

Redis

GemFire

各数据需求方

产品1

产品2

产品3

……

产品N

项目实例

- 看看在不同行业大家都是怎么用的

Conviva

Video Player Agents

视频事件

消息控制

Gateway

几秒

实时

实时指标

实时处理栈

分，时，天，月

跑在Hadoop上

历史

历史指标

离线处理栈

存储层

决策

决策

决策

决策

TV

PC

PAD

PHONE

good enough?

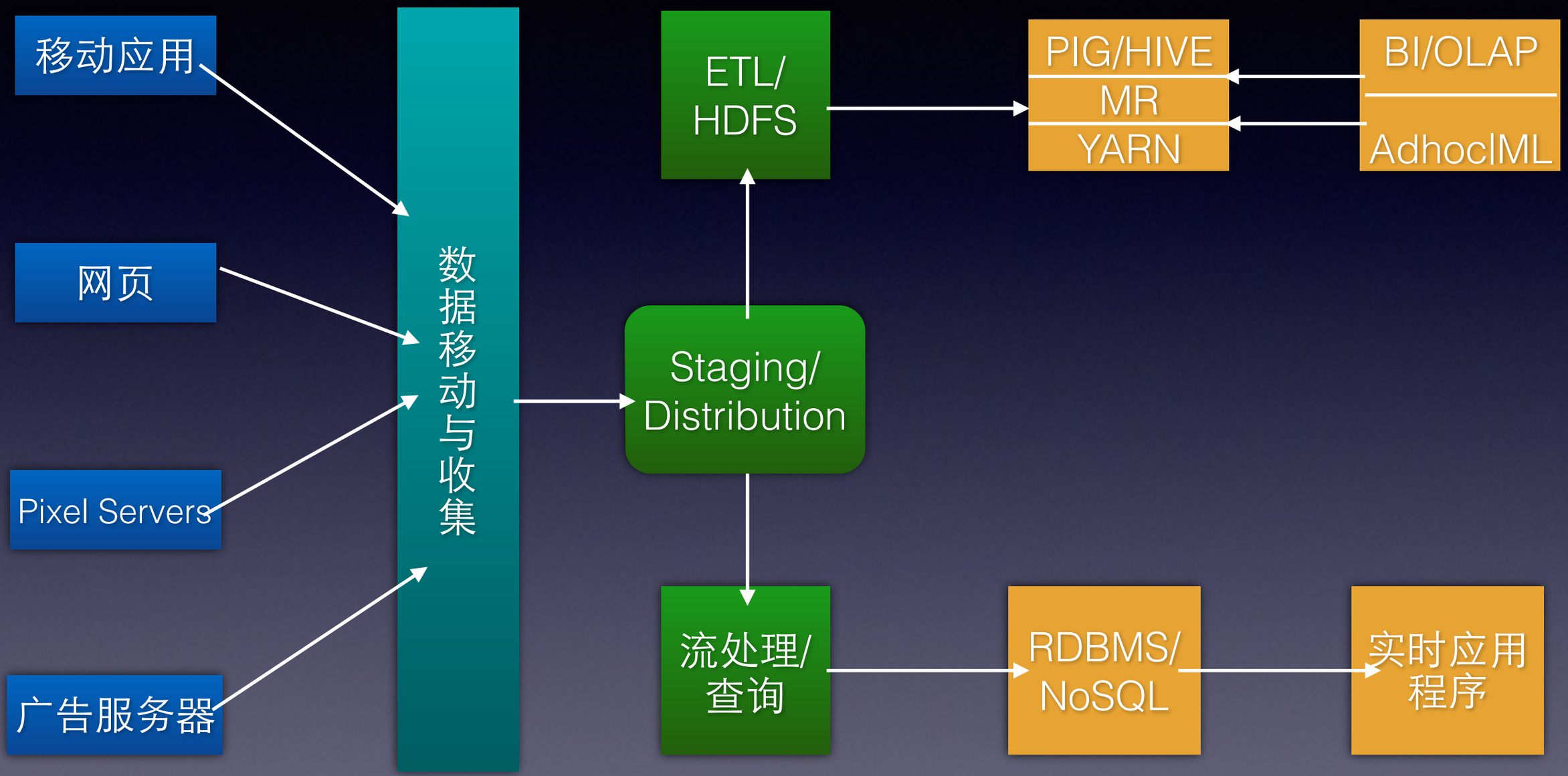
切记！！！

- 看起来很美好！！！BUT，细节是魔鬼！！！

- NEXT ——> 全部由Spark(& Spark Streaming)来完成

Yahoo!

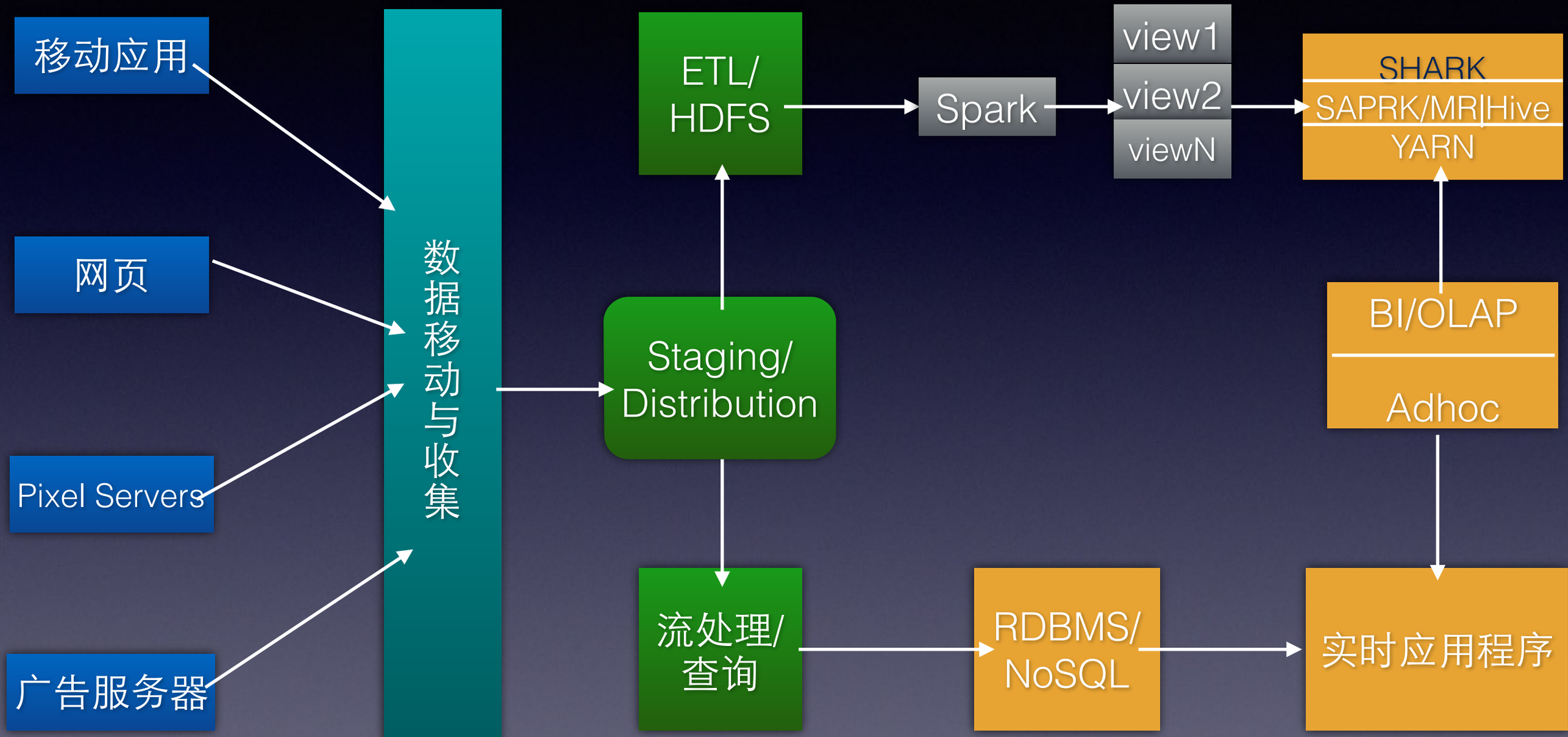
批处理



流处理

- 慢！慢！慢！　解决方案你应该已经想到~

批处理



流处理

- 很自然的替换/合作!

- 谢谢！ ~~~~~~~~~~~~或许，还需要多说几句

- Hadoop
- SPARK

小象学院
CHINA HADOOP

流处理

- Storm
- Spark Streaming
- S4

DB

- Cassandra
- HBase
- MongoDB
- Terrastore
- Redis
- SSDB
- MySQL

SQL on Hadoop(Spark)

- Hive
- Shark(Catalyst)
- Impala

日志(数据)收集

- Sqoop
- Flume
- Chukwa
- Kafka
- DataX
- Dbsync
- TimeTunnel

ML

- Mahout
- MLlib

Others

- Zookeeper
- Oozie
- Hue

- 没有最好的架构，只有最合适的架构！！！

谢 谢

欢迎关注我的公众微信号，经常会聊*Spark*相关的话题

ChinaScala

小象学院 CHINA HADOOP