大数据框架

1数据源

互联网、物联网、企业数据

2数据收集

把数据收集到大数据平台，ETL数据清洗;把不同服务器上面的数据收集到大数据平台

3数据存储

sql nosql

4资源管理

很好的协调计算资源、存储资源、网络资源，去处理数据；yarn（分布式资源管理系统，资源隔离和调度） 共享资源

５计算框架

批处理

交互式分析

类似使用sql查询数据库，但是数据量特别大

流处理

要求处理快速；无效点击检测

６数据挖掘

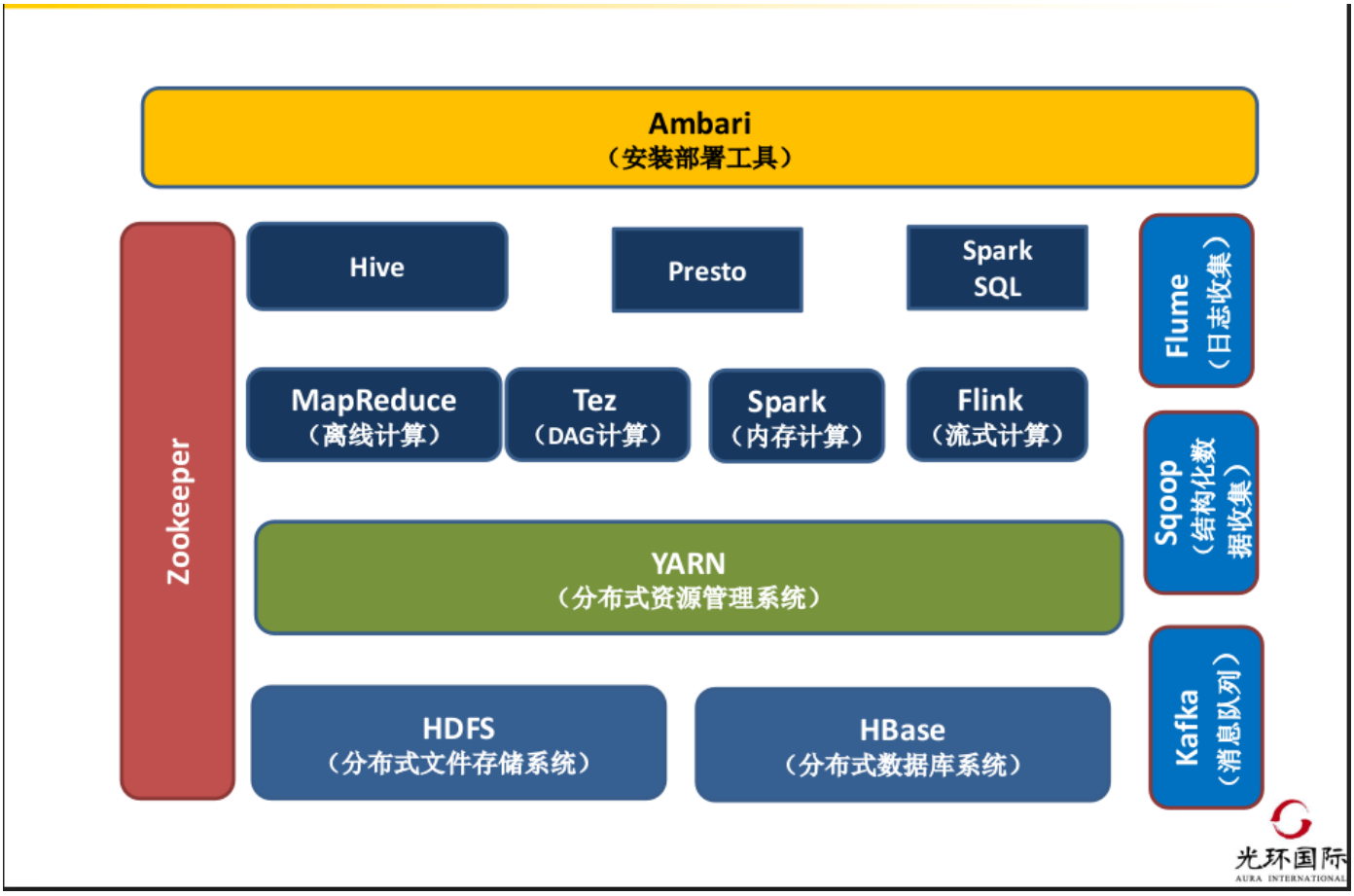
对数据进行分析

７数据可视化

数据展示

Hadoop生态系统

特点：1、源代码开源　2、社区活跃、参与者众多3、涉及分布式存储和计算的方方面面4、得到业界认可



Flume(非结构化数据收集)

Cloudera开源的日志收集系统

用于非结构化数据收集

Sqoop(结构化数据收集)

SQL-to-Hadoop 将传统关系型数据库中的数据传输到Hadoop中

HDFS(分布式文件系统)

源自于Google的GFS论文

特点：

1、良好的扩展性 2、高容错性 3、适合PB级以上海量数据的存储

基本原理

1、将文件切分成等大的数据块，存储到多台机器上；划分block,是主从结构（namenode是管理角色，datanode是工作者)

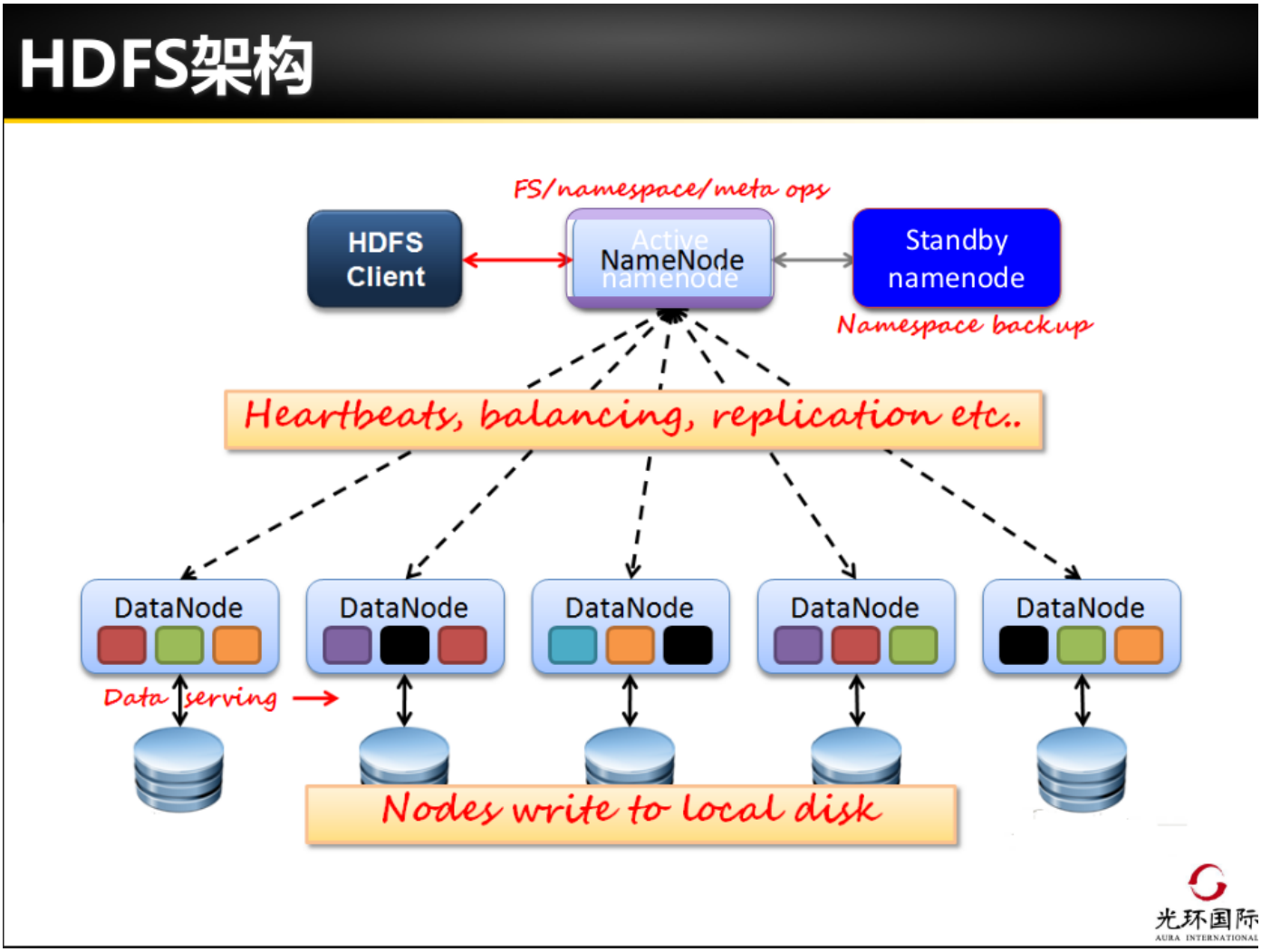
namenode 记录文件的目录树(由用户写进去的)；文件和block的映射的关系；block被存到了那些机器上（datanode汇报给namenode);记录datanode的状态（datanode心跳)

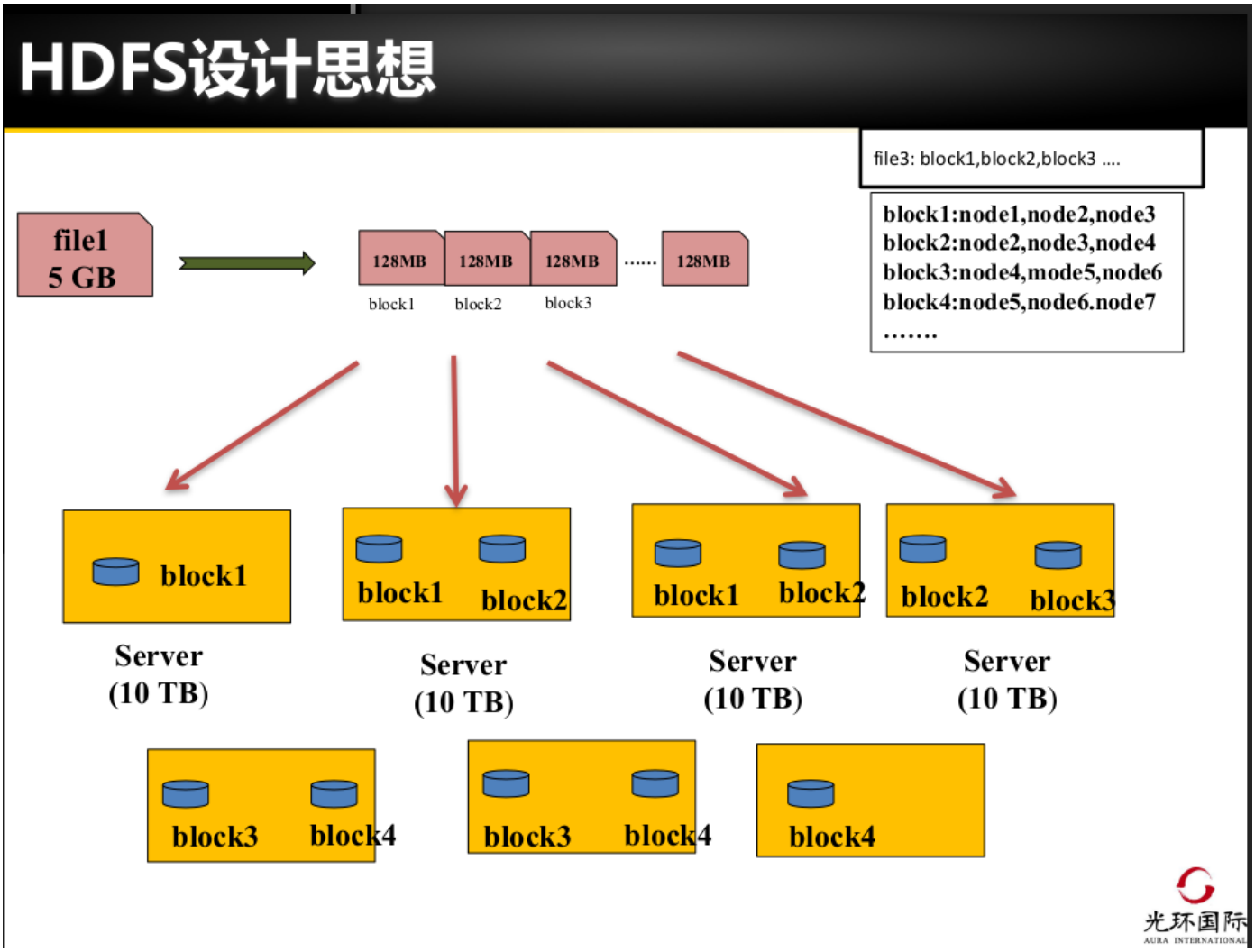
2、将数据切分、容错、负载均衡等功能透明化

3、可将HDFS看成一个容量巨大、具有高容错性的磁盘

hdfs不会将存储的文件映射到linux系统上；.meta是校验文件

小于128m的文件，block的大小就是文件的真实的大小





YARN(资源管理系统)

负责集群的资源管理和调度

使得多种计算框架可以运行在一个集群中

MapReduce(分布式计算框架)

分而治之

特点:

1、良好的扩展性

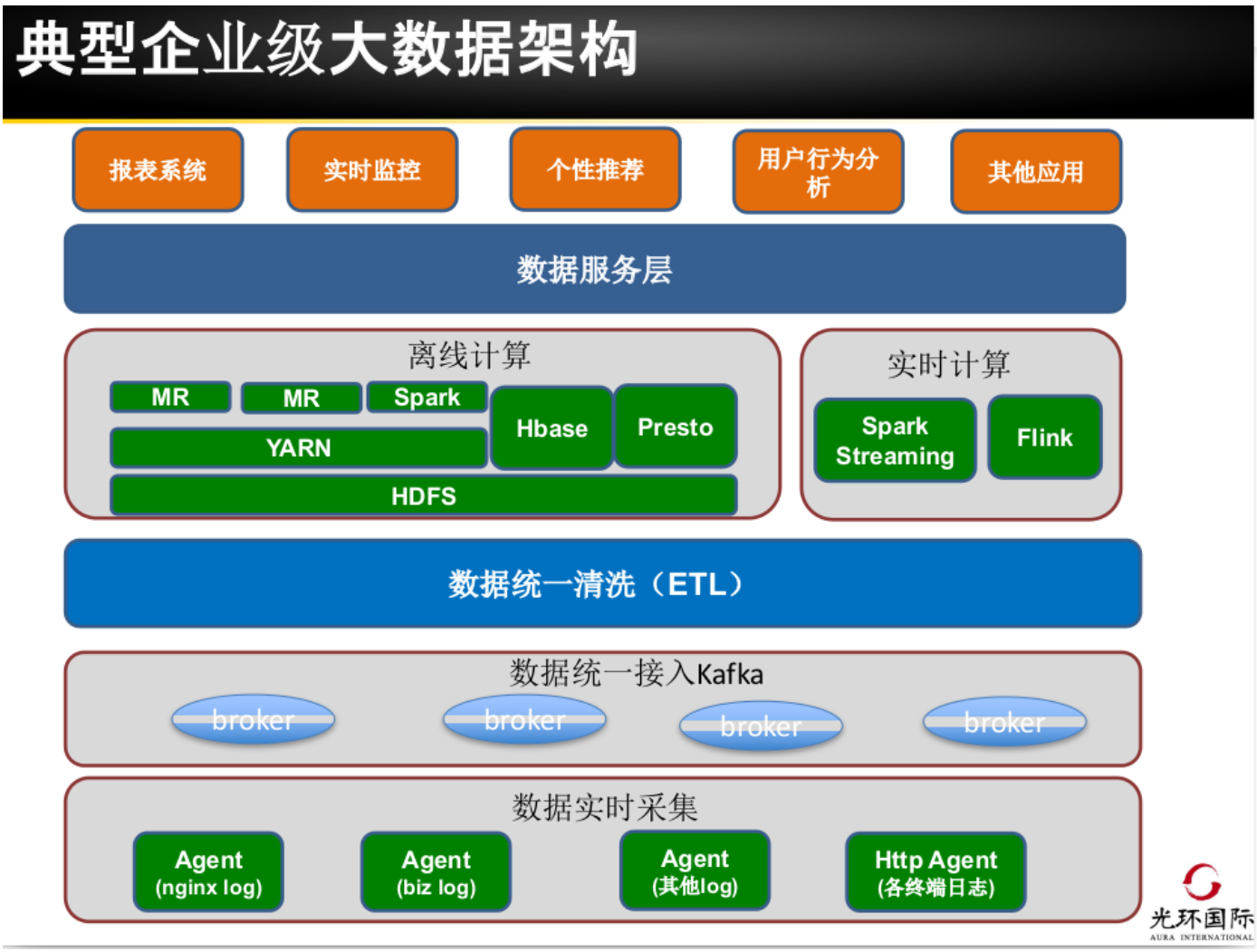
2、高容错性

3、适合PB级以上的海量数据的离线处理

Hive(基于MR的数据仓库)

是将sql翻译为MapReduce的库

企业大数据架构



Hadoop运行模式

本地模式

一个节点，不会启动任何服务

伪分布式模式

一个节点，所有服务都运行在该节点上；节点就是服务器

分布模式

多于一个节点

Type *Markdown* and LaTeX: α^2