1. Hadoop生态系统组件，源码

运维部署工具：

集群规模

1. Hive SQL, Spark SQL

Flume -> Kafka -> Spark Streaming -> Hive -> Sqoop -> MySQL

* Storm ->

Hadoop HDFS

系统运行过程中，哪些模块经常出问题？怎么解决，数据丢了怎么办

Spark程序的运行模式： YARN-Cluster，单独部署

RPC：

hive.skew

1. Hadoop，大规模集群瓶颈点

HDFS和YARN瓶颈点，NameNode Rpc ，YARN 调度等核心模块怎么去调试

依赖组件的瓶颈点

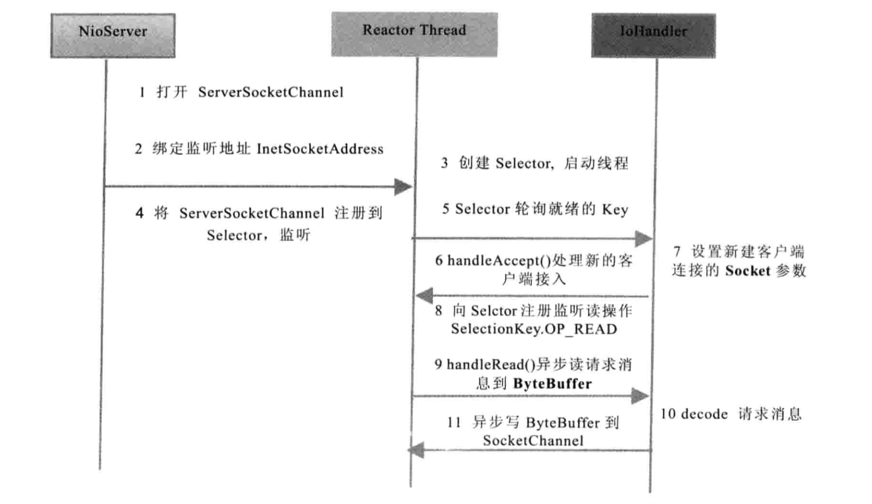
1. HDFS Block的放置策略，本地节点，机架等，MR本地性的理解
2. 数据仓库： Kylin: Hive及HBase

Hive数据仓库

Kylin

1. Hadoop生态系统的RPC，网络通信类库，序列化使用什么

JAVA NIO:非阻塞I/O



多路复用器Selector可以同时轮询多个Channel

Spark RPC:

Akka：1.6之前

Netty：

为了解决大块数据(如Shuffle)的传输问题

1. Hadoop 3.0的新特性

分布式调度

EC： 小文件，冷数据的处理方式

新资源的，IO DISK， Cgroup内存隔离

该同事从事Hadoop相关工作时间不长，在项目中基本上是使用Hive SQL来做数据分析，对Hadoop生态系统中的各软件（Kafka,Hbase,Spark 等）底层框架了解不是很多。对Hadoop集群的基本运维的能力欠缺，没有进行实际的部署及运维。

Ps.简历中还有错别字