# 10月17

1. 上午办理入职相关事宜，配置新电脑的环境。
2. 下午和魏建英对接智能运维业务。了解需要哪些知识储备。

# 10月18

1. 办理和签署入职协议。
2. 参加小组会议
3. 要学习spark，hive，kafka相关知识。
4. 先学习Spark相关技术：Spark基本原理介绍，与Hadoop的差异，核心模块介绍。快速开发案例。

# 10月19日

1. 参加内部评审会议
2. 学习spark基本原理
3. 阅读智能运维方案文档

# 10月20

1. 开会讨论智能运维整体框架
2. 阅读理解顾成龙发的知识库与智能运维文档。
3. 学习spark核心模块：
   1. Driver（master）
   2. Executor（slave）
   3. RDD : 弹性分布式数据集
      1. ⚫ 存储的弹性：内存与磁盘的自动切换；（当内存不够的时候可以存储在磁盘）
      2. ⚫ 容错的弹性：数据丢失可以自动恢复；
      3. （记录了读取那些文件和怎么划分的数据，所以数据丢失的时候可以重新读取）
      4. ⚫ 计算的弹性：计算出错重试机制；（计算出错可以从头开始计算）
      5. ⚫ 分片的弹性：可根据需要重新分片。（当Executor变化的时候可以同步改变任务和数据的划分）
      6. ➢ 分布式：数据存储在大数据集群不同节点上
      7. ➢ 数据集：RDD 封装了计算逻辑，并不保存数据
      8. ➢ 数据抽象：RDD 是一个抽象类，需要子类具体实现
      9. ➢ 不可变：RDD 封装了计算逻辑，是不可以改变的，想要改变，只能产生新的 RDD，在新的 RDD 里面封装计算逻辑
      10. ➢ 可分区、并行计算
      11. 5大核心属性
      12. ➢ 分区列表
      13. 分区列表，用于执行任务时并行计算，是实现分布式计算的重要属性。
      15. ➢ 分区计算函数
      16. 使用分区函数对每一个分区进行计算
      17. ➢ RDD 之间的依赖关系
      18. RDD 是计算模型的封装（装饰者模式），当需求中需要将多个计算模型进行组合时，就需要将多个 RDD 立依赖关系建
      19. ➢ 分区器（可选）
      20. 当数据为 KV 类型数据时，可以通过设定分区器自定义数据的分区
      21. 如何将数据分为不同的分区列表
      22. ➢ 首选位置（可选）
      23. 判断计算任务发送给那个节点是最优的
      24. （移动数据不如移动计算）
      25. 执行原理
      26. 数据处理过程中需要计算资源（内存 & CPU）和计算模型（逻辑）。执行时，需要将计算资源和计算模型进行协调和整合。
4. 开会讨论数据存储方案

# 10月21日

1. 开会讨论智能运维存储方案，了解知识库流程和业务。
2. 学习Spark
3. 跟hr核对入职合同，并邮寄给公司。