企业现行系统

分布式数据存储与云计算

技术预研报告

第一部分：HDFS、Yarn 研究报告

|  |  |  |
| --- | --- | --- |
| 文件状态：  [√] 草稿  [ ] 正式发布  [ ] 正在修改 | 文件标识： | AISINO-REPO-HADOOP-REPORT-001 |
| 当前版本： | 0.1A |
| 作 者： | Zrz |
| 完成日期： | 2016-08-29 |

版 本 历 史

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 版本/状态 | 作者 | 参与者 | 起止日期 | 备注 |
| 草稿 | Zrz |  | 2016-08-29 | 开始合并分析报告和预研日志 |
|  |  |  |  |  |
|  |  |  |  |  |

目 录

[0. 文档介绍 4](#_Toc7493470)

[摘要 4](#_Toc7493471)

[读者对象 4](#_Toc7493472)

[参考文档 4](#_Toc7493473)

[术语与缩写解释 4](#_Toc7493474)

[1. 背景介绍 5](#_Toc7493475)

[2. 技术预研目标 5](#_Toc7493476)

[3. 技术预研取得的工作成果 5](#_Toc7493477)

[4. hadoop的研究报告 5](#_Toc7493478)

[5. 商业智能的研究报告 5](#_Toc7493479)

# 0. 文档介绍

## 0.1 文档目的

首先需要明确的一点是：本阶段预研工作无意去回答“我们现阶段是不是需要建立大数据平台”这个问题，原因是：虽然我没有拿到业务部门围绕企业发展战略建立的业务增量模型，但无论从业内50TB这个起跳线还是以现行项目的商业量级来考量，这个问题似乎都并不紧迫；然而，考虑到战略层面的未雨绸缪和战术角度知己知彼的需要，这个主题的预研工作似有其必要性和现实意义。

整个预研将分三个部分，围绕分布式文件系统集群环境的搭建和运行表现、云计算、商业分析三个专题去揭示大数据平台的搭建成本与实际收益平衡的黄金点；同时枚举业界常规的技术手段以呈现这条技术路线上各种可能的技术门槛，为团队人力技术规划提供必要的参考；最后，希望为开发人员乃至需求人员在应对此领域需求时给予相对理性的出发点和必要佐证。

## 0.2 文档范围

## 0.3 读者对象

|  |  |
| --- | --- |
| **阅读对象** | **阅读建议** |
| 决策 |  |
| 需求 |  |
| 编码 |  |
| 评审 |  |

## 0.4 参考文档

## 0.5 术语与缩写解释

|  |  |
| --- | --- |
| **缩写、术语** | **解 释** |
| HDFS | 分布式文件系统 |
| Yarn | 资源管理器 |
| MapReduce | 云计算框架 |
| Sqoop | 关系型数据库与HDFS相互转化的工具 |
| Flume | 分布式日志聚合系统 |
| Storm |  |
| OLAP |  |
| Mahout |  |
| Spark |  |
| Zookeeper |  |
| Kafka |  |

`

# 1. 背景介绍

***提示：****说明为什么要开展技术预研 。*

# 2. 技术预研目标

***提示：****说明本次技术预研的主要内容与目标（必须是可以验证的）。*

# 3. 技术预研取得的工作成果

***提示：****说明本次技术预研取得的工作成果（程序、文档、数据等）以及时间。*

# 4.研究报告

***提示：***

*（1）解释重要的术语、基本概念；*

*（2）论述思想方法、技术方案、技术细节*

## 第一部分：HDFS、Yarn 研究报告

## 第二部分： 云计算研究报告