**目录**

# 概述

## 系统建设目标

该系统是基于美邦现有的平台，利用开源的大数据hadoop、hive、zookeeper、sqoop等技术，完成复杂的数据分析、数据导入导出等功能，以满足美邦日益增长的大数据需求，减少系统的数据库压力。

## 文档作用

该文档主要对美邦大数据hive平台设计方面进行了概要描述，包括了系统运行环境、总体架构、模块的划分、模块的功能、模块的流程及其相互间的接口关系。

本文档的使用范围为参加该项目建设的业务人员及技术人员。

## 术语

本文档中使用到的专门术语及其定义如下：

* hadoop

Hadoop是一个开发和运行处理大规模数据的软件平台,是Apache的一个用java语言实现开源软件框架，实现在大量计算机组成的集群中对海量数据进行分布式计算

* Hive

hive是基于Hadoop的一个数据仓库工具，可以将结构化的数据文件映射为一张数据库表，并提供简单的sql查询功能，可以将sql语句转换为MapReduce任务进行运行。

* Sqoop

Sqoop是一个用来将Hadoop和关系型数据库中的数据相互转移的工具，可以将一个关系型数据库（例如 ： MySQL ,Oracle ,Postgres等）中的数据导进到Hadoop的HDFS中，也可以将HDFS的数据导进到关系型数据库中。

* Hdfs

HDFS（Hadoop Distributed File System）是Hadoop项目的核心子项目，是分布式计算中数据存储管理的基础，Hadoop分布式文件系统(HDFS)被设计成适合运行在通用硬件(commodity hardware)上的分布式文件系统。

## 参考资料

* 《hadoop权威指南》

# 系统概述

## 需求综述

### 业务需求简述

基于ERP现有平台，结合hadoop大数据的相关技术，系统主要实现数据导入导出、查询分析等功能。

### 技术需求简述

系统根据需求在指定时间跑批作业，从老ERP系统获取源数据，经过hadoop大数据平台进行分析后，将跑批处理结果输出到新ERP系统。

Hadoop大数据平台监控页面可以监控到各个job作业的跑批结果，根据日志可以定位到错误信息、异常信息。

### 接口需求简述

Hadoop平台中的sqoop实现数据导入导出，通过JDBC连接新老ERP。

美邦hadoop大数据跑批作业平台，通过MbInitHiveHandler初始化hive表数据以及启动各JOB作业，批处理作业并发执行(根据业务需求也可以串行)。

# 系统总体设计

本章中对系统的总体架构进行描述。首先分析系统的层次结构，然后说明系统的基础架构，并对系统的各个package进行了说明，最后对系统中异常处理进行了说明。

## 系统交互图

美邦老ERP系统：为hadoop大数据各种业务作业分析提供数据源

Oracle：存放业务和交易数据。关系型数据库

Hive: 数据仓库工具，数据分析

Sqoop: 数据导入导出工具

## 层次结构

### 门面层

门面层主要包装业务处理层服务，向外提供http/WEB Service接口。在门面层中统一处理业务处理层接口的异常，包装成统一的返回结果。

### 业务处理层

业务处理层的核心是hive批处理作业，进行各种复杂的业务分析。

### 数据持久层

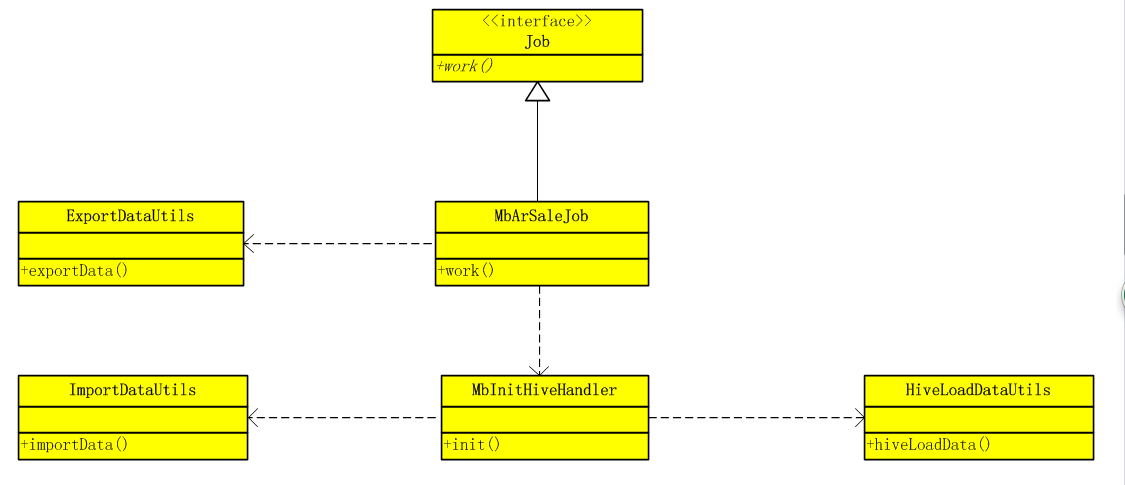
数据持久层主要完成数据的持久化操作，供业务处理层调用，使用IBatis实现。

# **批处理作业应用**

## 门店销售数据

### Java类及接口设计

#### 类图

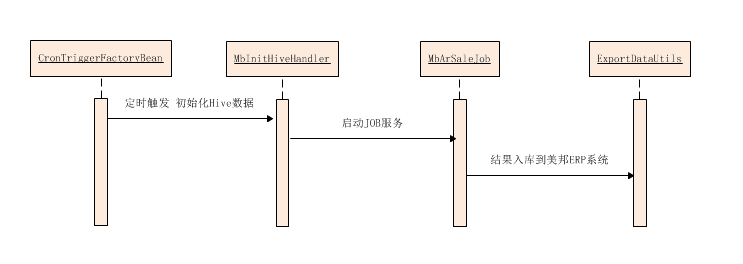


包：com.mb.saas.bi.job

接口名：Job,批处理作业统一接口

实现类：MbArSaleJob，统计门店销售数据

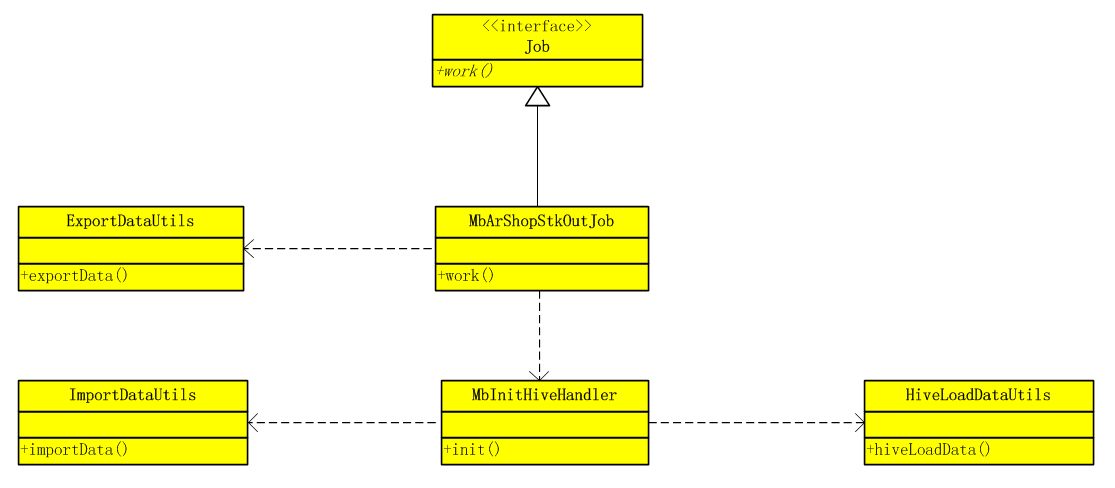
#### 时序图



## 门店出库商品

### Java类及接口设计

#### 类图

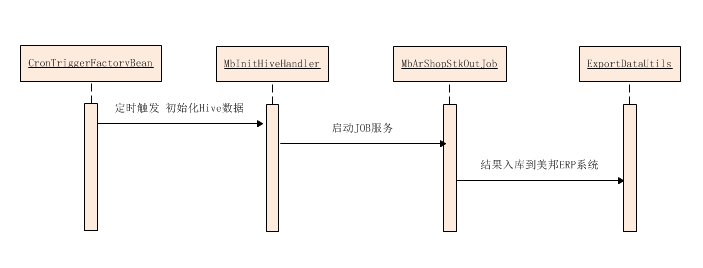


包：com.mb.saas.bi.job

接口名：Job,批处理作业统一接口

实现类：MbArShopStkOutJob，统计门店出库商品信息

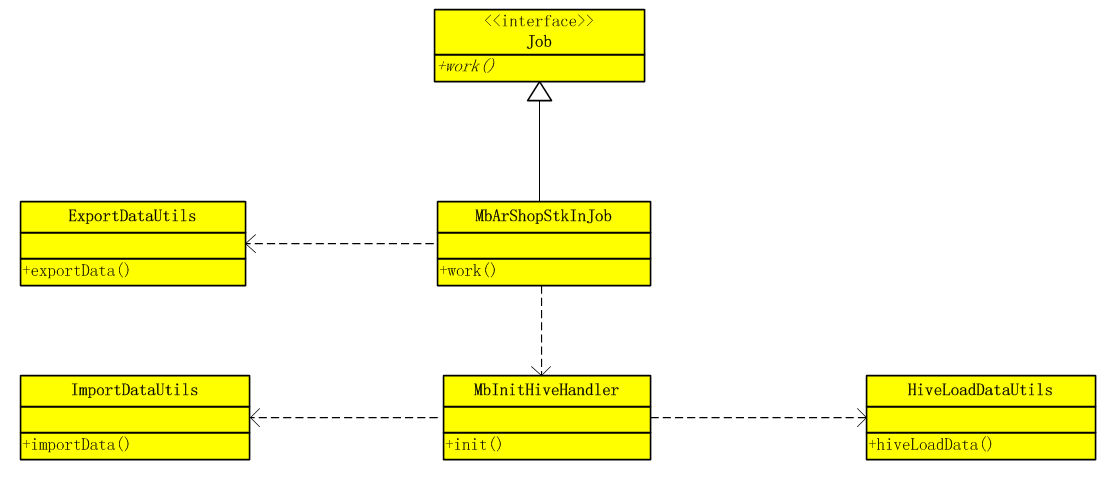
#### 时序图



## 门店入库商品

### Java类及接口设计

#### 类图

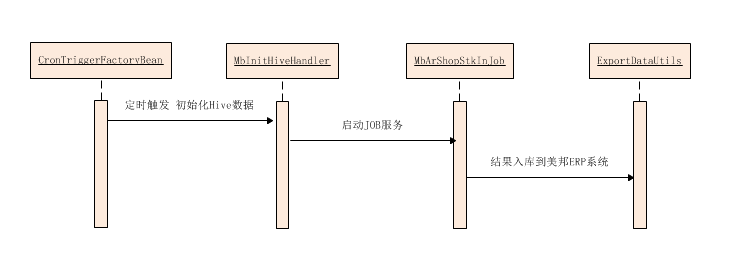


包：com.mb.saas.bi.job

接口名：Job,批处理作业统一接口

实现类：MbArShopStkInJob，统计门店入库商品信息

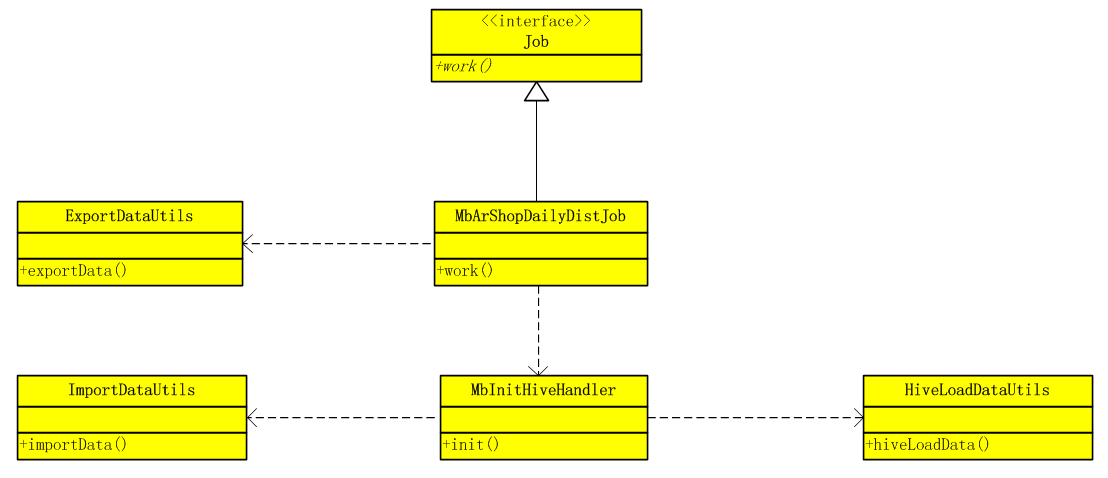
#### 时序图



## 门店订货每日发货数据

### Java类及接口设计

#### 类图

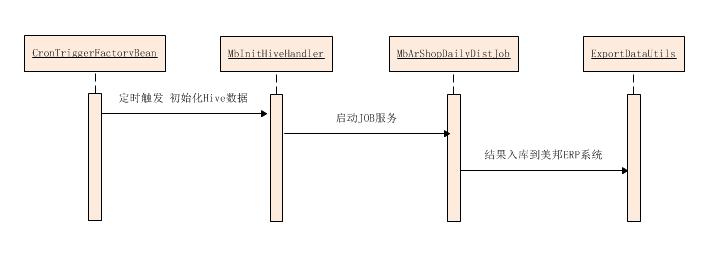


包：com.mb.saas.bi.job

接口名：Job,批处理作业统一接口

实现类：MbArShopDailyDistJob，统计门店每日发货数据

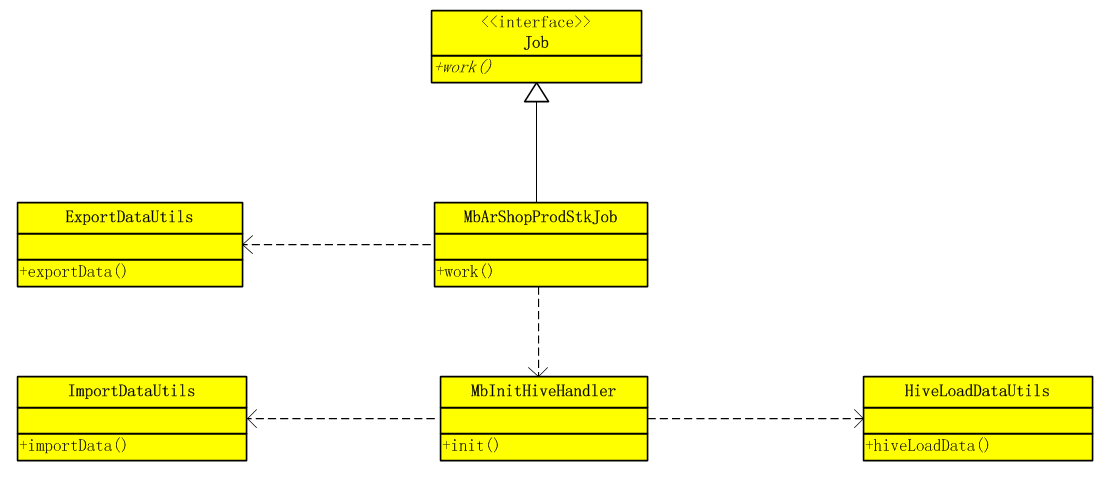
#### 时序图



## 门店库存

### Java类及接口设计

#### 类图

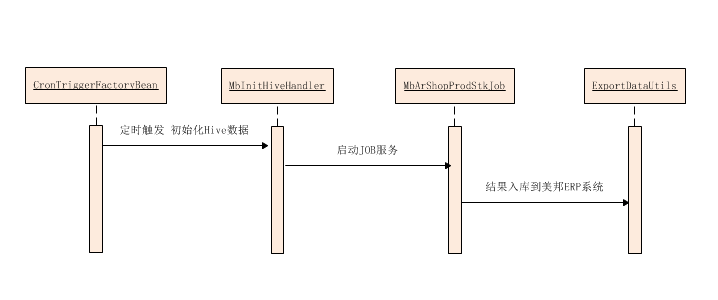


包：com.mb.saas.bi.job

接口名：Job,批处理作业统一接口

实现类：MbArShopProdStkJob，统计门店库存信息

#### 时序图



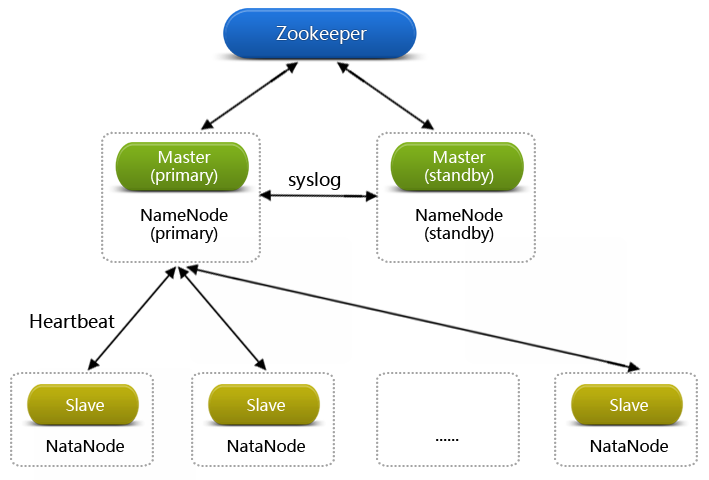
# **物理架构设计(集群部署图)**

### 集群部署

这里采用 5台虚拟机，内存是8G，硬盘是200GB

如图所示：一个主节点，一个备份节点，三个子节点;Hive部署在子节点上(slave1)，减少主节点的负载压力;zookeeper集群负责hadoop集群的分布式协调管理

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ip地址 | 主机名 | NameNode | JournalNode | DataNode | Zookeeper | Hive |
| 10.100.22.251 | master | 是 | 是 | 否 | 是 | 否 |
| 10.100.22.176 | backup | 是 | 是 | 否 | 是 | 否 |
| 10.100.22.218 | Slave1 | 否 | 是 | 是 | 是 | 是 |
| 10.100.22.219 | Slave2 | 否 | 否 | 是 | 是 | 否 |
| 10.100.22.242 | Slave3 | 否 | 否 | 是 | 是 | 否 |



### 集群部署注意事项

1. master节点的硬件尽量配置好点，而dataNode的硬件配置需要保持一致；
2. 当三个子节点不能满足生产环境的需要时，可申请增加节点
3. Hadoop集群的运维监控界面，可以监控集群运行情况, <http://10.100.22.251:8088/cluster/> 如图

