讨论课 01

1352839
饶伊文
项目 git 链接：https://github.com/BrunoQin/Open-Reuse

一、讨论课内容要求：

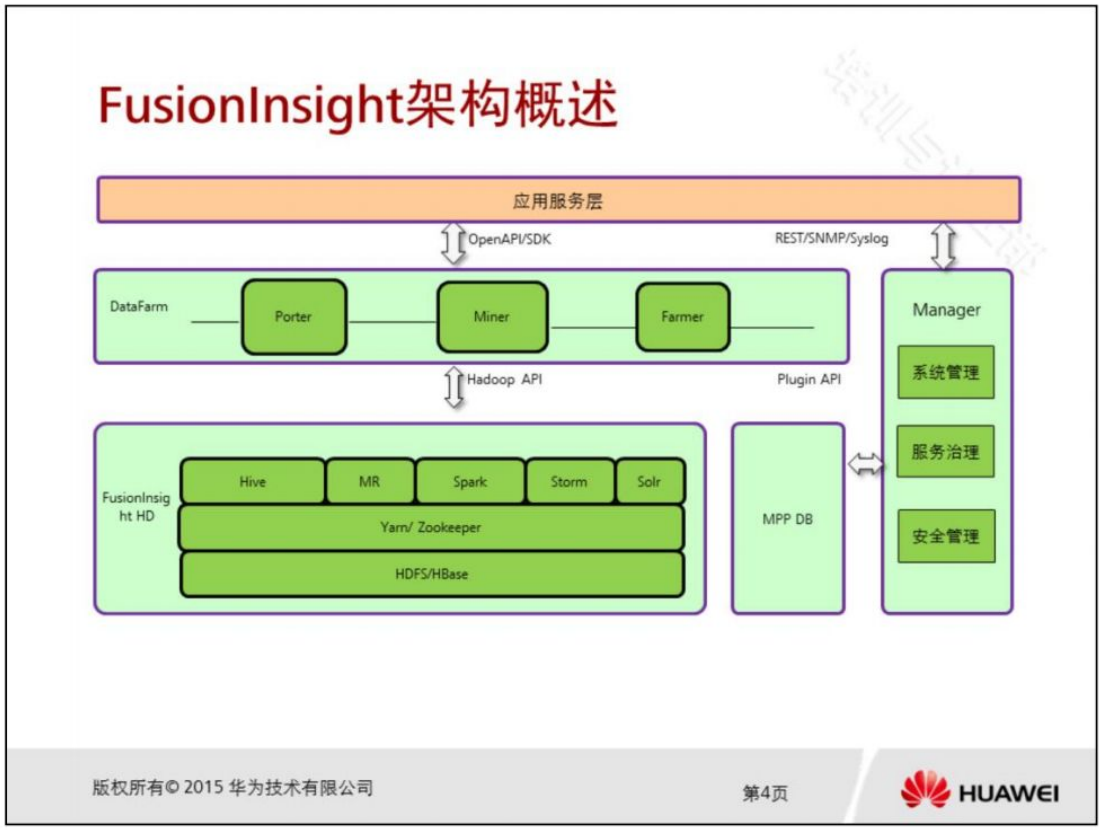
参考业界架构，考虑程序的扩展：

分布式：高可用性，高吞吐量
数据储存，分区，一致性，缓存
负载均衡
系统监控
ID 分配
通信可靠高效
协议
消息队列
垃圾消息过滤
安全
。。。

二、内容分析与业界参考方案：

本部分对华为的分布式大数据处理系统 FusionInsight 进行理解分析，通过此案例可以看出分布式系统的高可用性与高吞吐量

1、架构概述



华为 FusionInsight 是一个分布式数据处理系统，对外提供大容量的数据存储、查询和分析能力，可解决各大企业的以下需求。 FusionInsight 的 Hadoop 层提供大数据处理环境，基于社区开源软件增强，按照场景选择业界最佳实践。

FusionInsight 的 DataFarm 层提供支撑端到端数据洞察，构建数据到信息到知识到智慧的数据供应链，其中包括相对独立的数据集成服务 Porter，数据挖掘服务 Miner 和数据服务框架 Farmer。

FusionInsight Manager 是一个分布式系统管理框架，管理员可以从单一接入点操控分布式集群，包括系统管理（ OM/NTP/灾备）、数据安全管理和数据治理。

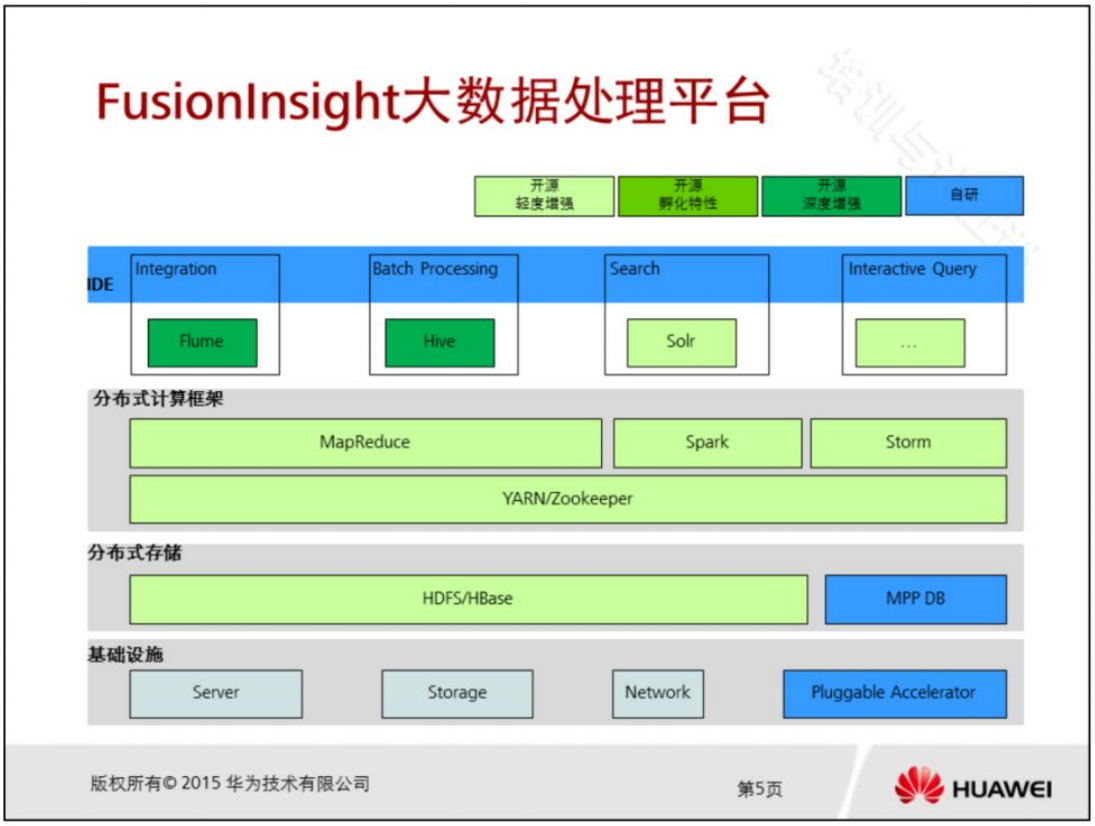
FusionInsight Stream：提供实时流处理平台。

FusionInsight Farmer：提供数据服务框架，大数据实时应用使能器，支撑企业快速开发基于大数据平台的应用。

FusionInsight Miner：提供数据挖掘服务集，基于分布式内存计算的数据分析平台。

FusionInsight MPPDB：提供相对独立部署的通用 MPP 数据库，用于性能较高的交互分析场景。

2、大数据处理平台



FusionInsight 用 100%开源的核心支持混合负载，从批量、交互查询、数据挖掘，到实时流和查询等各种场景，所有的组件都通过 Manager 提供的插件框架来按需安装。

基础设施：通用服务器、存储、网络设备和可插拔加速器。

分布式存储：

HDFS：分布式文件系统。

HBase： Hadoop 分布式数据库。

MPP DB：相对独立部署的通用 MPP 数据库，用于性能较高的交互分析场景。

分布式计算框架：

MapReduce：基于磁盘的离线分布式计算框架。

Spark：基于内存的迭代计算框架。

Storm：实时的、分布式，在线实时流处理计算系统。

Yarn：资源管理与调度系统。

Zookeeper：是一个开源文件应用程序接口，能使大型系统的分布式进程相互同步，这样所有提出请求的客户端就可以得到一致的数据，而且可以避免单点故障。

IDE： Integrated Development Environment，集成开发环境。

Integration：数据集成，例如： Flume 是一个分布式、可靠和高可用的海量日志聚合系统。

Batch Processing：批处理，例如： Hive 是建立在 Hadoop 上的数据仓库框架，提供类似 SQL 的 HQL 语言操作结构化数据。

Search：搜索，例如： Solr 提供全文搜索引擎。

三、参考资料：

华为网络技术大赛学习资料：

1.1 FusionInsight 系统概述