Cloudera 5.3.x文档学习笔记

harli

2015/2/4

Shanghai-pd

# Cloudera介绍

<http://www.cloudera.com/content/cloudera/en/documentation/core/latest/topics/introduction.html>

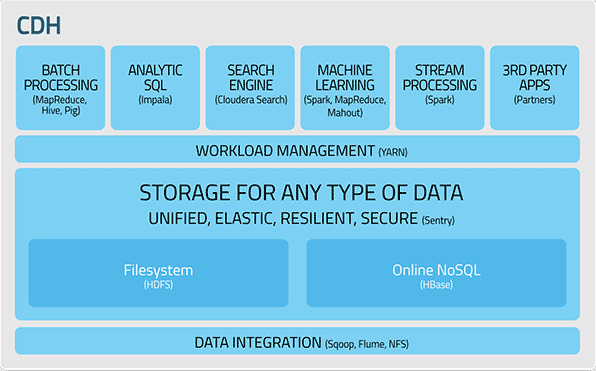
## CDH

CDH是Apache Hadoop及其相关项目的最完整的,经过测试的流行发布版本。CDH提供了Hadoop最核心的部分 —— 可扩展的存储以及分布式计算 ——并带有一个基于web的用户界面和至关重要的企业功能。CDH是apache许可开源的，并且是唯一一个提供了统一的批处理、交互式SQL和交互式search，以及基于角色（role）的访问控制的Hadoop解决方案。

CDH提供了：

1. 灵活性——存储任何类型的数据，并且可以通过各种不同的计算框架，包括批处理、交互式SQL,免费的文本搜索,机器学习和统计计算等对数据进行操作。
2. 集成 — 快速启动和运行一个完整的Hadoop平台,适用于广泛的硬件和软件解决方案。
3. 安全性：处理和控制敏感数据。
4. 可扩展性——允许广泛的应用和规模，通过扩展来满足您的需求。
5. 高可用性——关键的业务任务的执行是可信任的。
6. 兼容性——可以利用你现有的IT基础设施和投入。

## CDH的架构



## Cloudera Impala

Cloudera Impala提供了直接针对存储在HDFS或HBase上的Apache Hadoop数据的快速的、交互式的SQL查询。除了使用统一的存储平台，Impala也使用了相同的metadata，SQL语法（Hive SQL），ODBC driver，以及和Apache Hive一样的用户界面。提供了一个用于实时或面向批处理的查询的常见又统一的平台。

ClouderaImpala是另外一个提供大数据查询的可用工具。Impala不能替代建立在MapReduce基础上的批处理框架，如Hive。Hive和其他建立在MapReduce基础上的框架最适合用于长时间运行的批处理作业,如涉及到提取，转换和加载(ETL)类型的批处理工作。

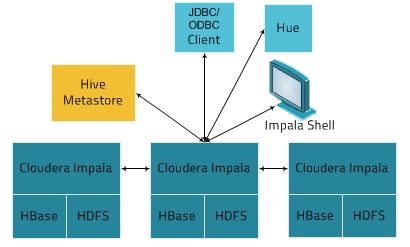
### Impala 的优点

Impala 提供了:

1. 数据科学家和分析师已经知道熟悉的SQL接口；
2. 在Apache Hadoop大数据上的交互式查询能力；
3. 在集群环境中进行分布式查询,方便扩展，并且可以使用廉价的商用硬件；
4. 不需要通过拷贝或导出/导入等步骤就可以在不同组件之间共享数据文件的能力；比如，可以使用Pig写数据，并用Impala读取数据，或者使用Impala写数据，然后使用Hive读取。
5. 在单个系统上进行大数据处理和分析，使顾客避免仅仅为了数据分析而进行的昂贵的建模和ETL。

### Cloudera Impala 如何工作在CDH上

下图描述了Impala如何部署在大规模的Cloudera环境中：



Impala解决方案由以下组件组成：

1. 客户端——包括Hue、ODBC客户端,JDBC客户机和Impala Shell这些实体，都可以与Impala进行交互。这些接口通常用于问题查询或完成连接Impala的管理任务。
2. Hive Metastore——存储Impala可用数据的信息。例如,metastore可以让Impala知道哪些数据库是可用的，以及这些数据库的结构是什么。当你通过Impala SQL语句创建、删除和修改Schema对象, 加载数据到tables中等等,相关的元数据修改就会通过专门的，在Impala 1.2 中引入的catalog服务，自动广播到所有Impala节点。
3. Cloudera Impala —— 运行在datanode上的进程,会协调和执行查询。每一个Impala示例可以可以接收、计划和协调来自Impala客户端的查询。查询会被分布到Impala的各个节点上,然后这些节点作为Workers,去执行并行的查询片段（fragments）。
4. HBase和HDFS -存储用于查询的数据。

使用Impala执行的查询，其处理步骤如下：

1. 用户应用程序通过提供标准化查询接口的ODBC、JDBC，发送SQL查询到Impala。用户应用程序可以连接到集群中任何一个impalad。这个impalad为查询的协调员。
2. Impala解析查询并进行分析,以确定哪些任务需要被跨集群（across the cluster）impalad实例优化。为了实现最佳效率而对执行进行计划。
3. 像HDFS和HBase这些服务，由本地impalad实例进行访问，来提供数据。
4. 每个impalad将返回数据到协调impalad，它会将结果发送到客户端。

### Impala的主要特性

Impala提供的支持：

1. Hive查询语言(HiveQL)最通用的sql - 92的特性，包括SELECT，joins，和aggregate函数。
2. HDFS和HBase存储,包括:
3. [**HDFS 文件格式**](http://www.cloudera.com/content/cloudera/en/documentation/core/latest/topics/impala_file_formats.html#file_formats):文本文件、SequenceFile RCFile,Avro文件, Parquet。
4. Compression codecs: Snappy, GZIP, Deflate, BZIP.
5. 常见的Hive的接口包括
6. [**JDBC driver**](http://www.cloudera.com/content/cloudera/en/documentation/core/latest/topics/impala_jdbc.html#impala_jdbc).
7. [**ODBC driver**](http://www.cloudera.com/content/cloudera/en/documentation/core/latest/topics/impala_odbc.html#impala_odbc).
8. Hue Beeswax 和新的Cloudera Impala 查询界面
9. Impala [**命令行接口**](http://www.cloudera.com/content/cloudera/en/documentation/core/latest/topics/impala_impala_shell.html#impala_shell).
10. [**Kerberos authentication**](http://www.cloudera.com/content/cloudera/en/documentation/core/latest/topics/impala_security.html#security).

## Cloudera Search

## Cloudera Search