Turing AI 万维图灵 | 大数据系列课程

大数据技术-第一章：Hadoop大数据概述

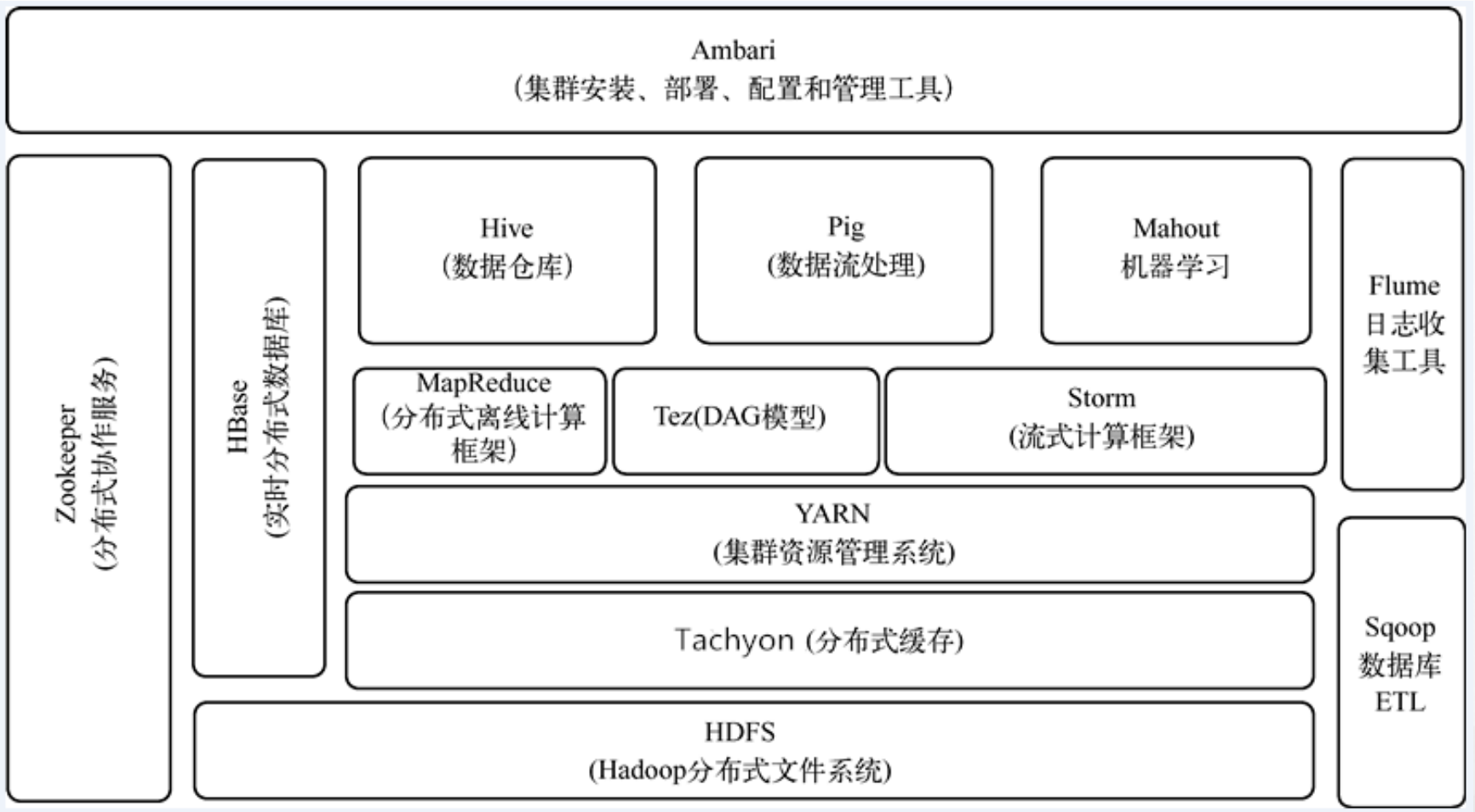
Hadoop生态系统简介

CONTENTS

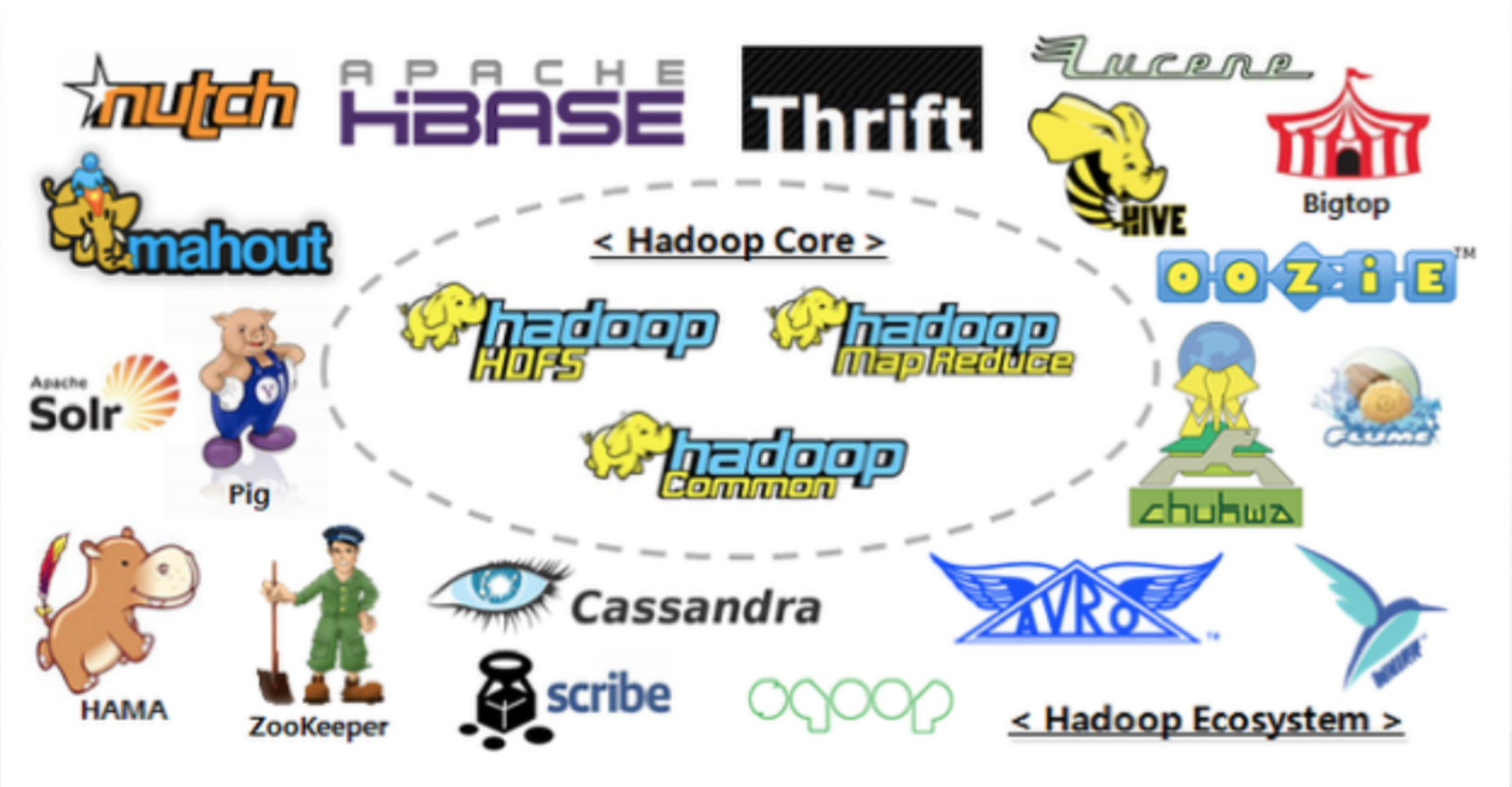
01

Hadoop生态系统

Hadoop生态系统


Ambari
(集群安装、部署、配置和管理工具)
Zookeeper
(分布式协作服务)
HBase
(实时分布式数据库)
Hive
(数据仓库)
Pig
(数据流处理)
Mahout
机器学习
Flume
日志收集工具
MapReduce
(分布式离线计算框架)
Tez(DAG模型)
Storm
(流式计算框架)
YARN
(集群资源管理系统)
Tachyon (分布式缓存)
Sqoop
数据库ETL
HDFS
(Hadoop分布式文件系统)

Hadoop生态系统


nutch
APACHE HBASE
Thrift
lucene
HIVE
Bigtop
mahout
< Hadoop Core >
OOZIE
hadoop HDFS
hadoop MapReduce
hadoop Common
Apache Solr
Pig
chukwa
FLUME
HAMA
ZooKeeper
Cassandra
AVRO
scribe
< Hadoop Ecosystem >

02

HBase

HBase – Hadoop Database，是一个高可靠性、高性能、面向列、可伸缩的分布式存储系统，利用HBase技术可在廉价PC Server上搭建起大规模结构化存储集群。

Row	Timestamp	Animal		Repair
		Type	Size	Cost
Enclosure1	12	Zebra	Medium	1000€
	11	Lion	Big	
Enclosure2	13	Monkey	Small	1500€

Region

Key

Column

Family

Cell

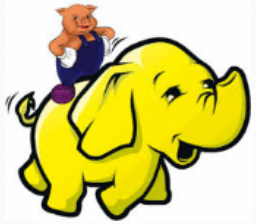
(Table, Row_Key, Family, Column, Timestamp) = Cell (Value)

03

数据访问

数据访问

Hive是建立在Hadoop 上的数据仓库基础构架。它提供了一系列的工具，可以用来进行数据提取转化加载（ETL），这是一种可以存储、查询和分析存储在 Hadoop 中的大规模数据的机制。



Pig是一个基于Hadoop的大规模数据分析平台，它提供的SQL-LIKE语言叫Pig Latin，该语言的编译器会把类SQL的数据分析请求转换为一系列经过优化处理的MapReduce运算。

共同点：

都是把代码转换为MapReduce任务；

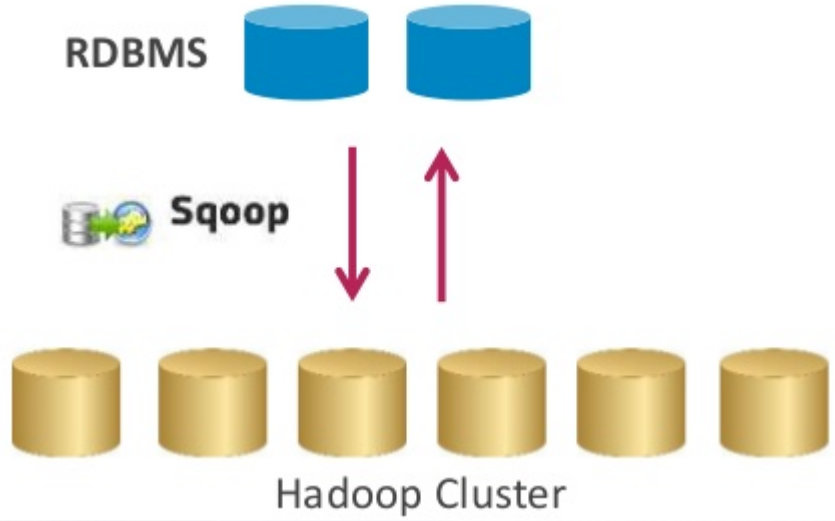
不同点：

Hive使用SQL、Pig使用pig Latin；

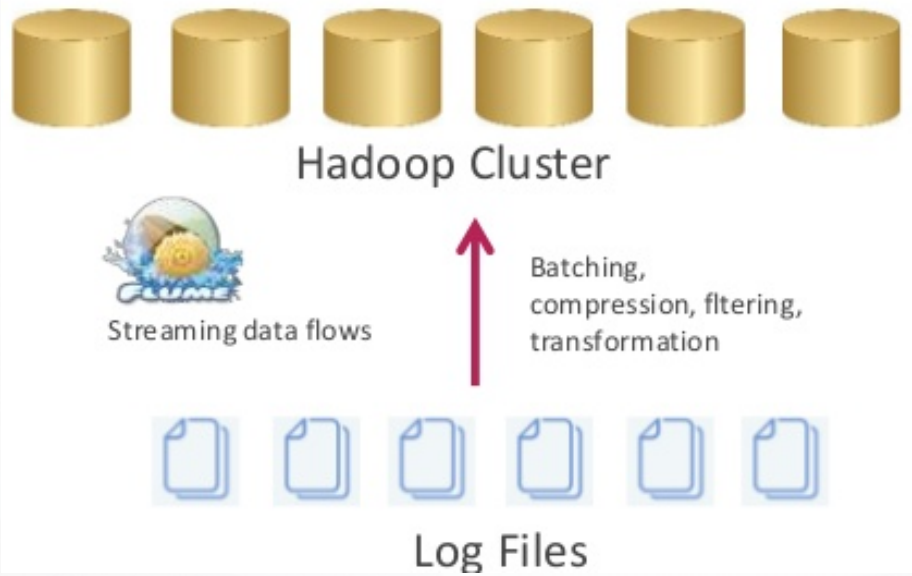
04

数据传输

数据传输

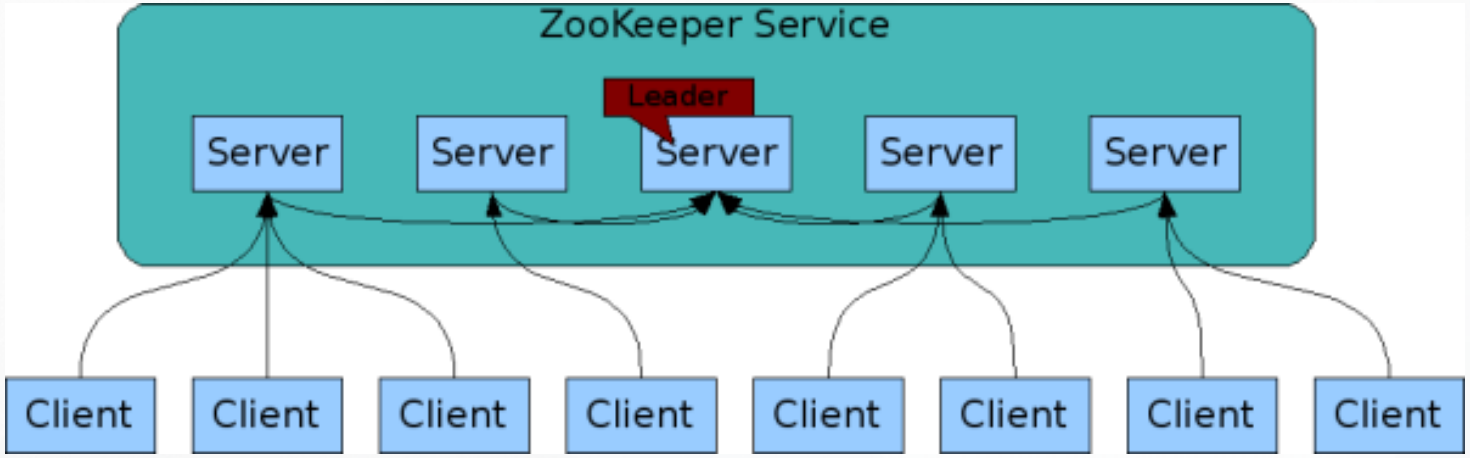


Sqoop是一款开源的工具，主要用于在Hadoop(Hive)与传统的数据库(mysql、postgresql...)间进行数据的传递，可以将一个关系型数据库中的数据导进到Hadoop的HDFS中，也可以将HDFS的数据导进到关系型数据库中。



Flume是Cloudera提供的一个高可用的，高可靠的，分布式的海量日志采集、聚合和传输的系统，Flume支持在日志系统中定制各类数据发送方，用于收集数据；同时，Flume提供对数据进行简单处理，并写到各种数据接受方（可定制）的能力。

05
管理



ZooKeeper是一个分布式的，开放源码的分布式应用程序协调服务，是Google的Chubby一个开源的实现，是Hadoop和Hbase的重要组件。它是一个为分布式应用提供一致性服务的软件，提供的功能包括：配置维护、域名服务、分布式同步、组服务等。

06
机器学习

Mahout 是 Apache Software Foundation（ASF） 旗下的一个开源项目，提供一些可扩展的机器学习领域经典算法的实现，旨在帮助开发人员更加方便快捷地创建智能应用程序。Mahout包含许多实现，包括聚类、分类、推荐过滤、频繁子项挖掘。此外，通过使用 Apache Hadoop 库，Mahout 可以有效地扩展到云中。

Turing AI 万维图灵 | 大数据系列课程

大数据

BIG DATA

智 / 能 / 科 / 技　　放 / 眼 / 未 / 来