

驾驭大数据

以Hadoop为核心的大数据开放平台建设

④ 孙利兵



讯飞开放平台

<http://www.xfyun.cn/>

大数据技术发展

cloudera manager



Apache Ambari

运维管理工具

- ④ 简易的集群部署功能
- ④ 服务配置管理
- ④ 集群状态监控



不断完善的生态系统

- ④ SQL数据操作 (Hive、Impala、Shark)
- ④ 脚本语言 (Pig)
- ④ ETL (Flume、Sqoop)
- ④ 内存计算&流计算 (Impala、Shark、Storm)
- ④ 工作流 (oozie)



架构不断优化

- ④ Yarn (第二代Mapreduce)
- ④ NameNode Federation



一头奔跑的大象，不断进化

- ④ Native lib
- ④ Checksum 机制
- ④ ShortCircuit Read

大数据技术发展

④ 大数据技术有哪些不足

- ④ 大数据技术本身百花齐放，如何用好每项技术是个难题
- ④ 大数据技术内部融合性不够
- ④ 大数据技术与其他传统技术的融合性不够

④ 我们缺少什么？

- ④ 缺乏一个能融合现有大数据技术的技术

技术领域如何解决大数据技术应用难的问题？

大数据基础技术的风向标

Apache Hadoop Ecosystem

Doug Cutting
Cloudera & Apache



The State of the Apache .pdf

大数据基础技术的风向标

- ④ The Ecosystem is the System
 - ④ Hadoop has become the kernel
 - ④ of the distributed operating system for Big Data
 - ④ a de-facto industry standard
- ④ No one uses the kernel alone
- ④ A collection of projects at Apache
- ④ Avro support across components



以Hadoop为核心，融合其他技术的平台系统
Avro是实现融合的关键技术

Cloudera在做-Hadoop应用体验

cloudera

Ask Bigger Questions

Hadoop, made easy.

Use Hadoop

Query Apache Hive and Cloudera Impala, search and customize Apache Solr, browse and manipulate files and directories in the Hadoop Distributed File System (HDFS), create and run Apache Pig scripts, visually manage Apache Oozie workflow/coordinator /bundle applications, create, submit and browse MapReduce jobs...



Hue


Administer Hadoop


Setup and monitor the health of the cluster, start and stop services like HDFS, Job Tracker, update and deploy configurations, search logs, perform audits, analyse performance graphs and metrics...






























Cloudera Manager

Cloudera在做-Hadoop开发体验

**cdk**
Cloudera Development Kit
Last updated a day ago

cdk

- kite-data-core
- kite-data-crunch
- kite-data-hbase
- kite-data-hcatalog
- kite-data-hive
- kite-data-mapreduce
- kite-data-spark
- kite-hadoop-compatibility
- kite-maven-plugin
- kite-morphlines-avro
- kite-morphlines-core
- kite-morphlines-hadoop-core
- kite-morphlines-hadoop-parquet-avro
- kite-morphlines-hadoop-rcfile
- kite-morphlines-hadoop-sequencefile
- kite-morphlines-json
- kite-morphlines-maxmind
- kite-morphlines-metrics-servlets
- kite-morphlines-protobuf
- kite-morphlines-saxon
- kite-morphlines-solr-cell
- kite-morphlines-solr-core
- kite-morphlines-tika-core
- kite-morphlines-tika-decompress
- kite-morphlines-twitter
- kite-morphlines-useragent
- kite-tools

Java ★ 137 ⓘ 57

Develop With CDK



Maple

讯飞如何应对这个技术挑战？

讯飞大数据开放平台



Maple
大数据开放平台

- 以数据导向为理念
- 以Hadoop为核心
- 融合优秀技术
- 因地制宜的使用技术
- 提升大数据用户体验

讯飞大数据开放平台的构成

大数据全新应用体验Evolution



Maple-
SDK

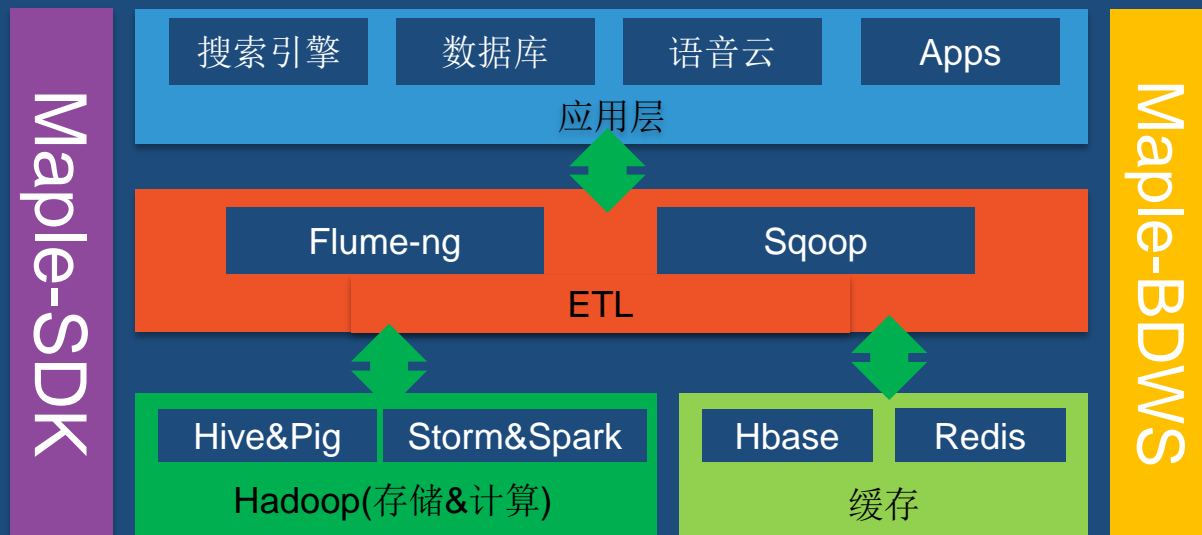


基础集群



Maple-
BDWS

讯飞大数据开放平台-架构图



大数据开放平台的门户



Maple-
BDWS

大数据工作站-Maple-BDWS

④ 功能

- ④ 代码托管
- ④ 编译部署
- ④ 工作流设计
- ④ 任务调度
- ④ 数据&任务信息浏览

④ 特点

- ④ 多个集群管理
- ④ 多版本集群兼容
- ④ 支持多项目管理
- ④ 在线编译部署 (One button to use)

大数据工作站- Maple-BDWS

Maple



Home



Oozie



Hive



HDFS



IPython



Webshell



Demo / maple-demo

🕒 Ibsun created on Mar 31, 2014

OPERATIONS



Clone ▾

★ Starred (5)

+ Follow(3)

Overview

Source Code

Build

🚩 Workflows ▾

OldWorkFlow

⚙️ Settings



master ▾

pull 📶

maple-demo /

HashCode

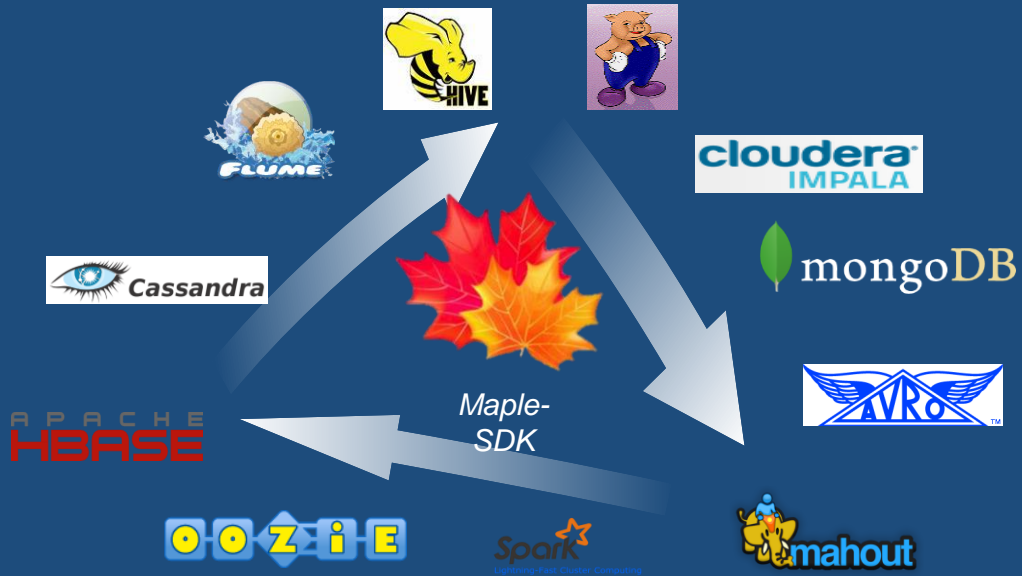
📁 bin	mv the deploy code from build.sh	6 months ago
📁 demo-data	rename txt in deme-data	6 months ago
📁 demo-dist	use demo-data for the test data dir.	6 months ago
📁 demo-mr	add GenerateMRData	6 months ago
📁 demo-wf	modify the hadoop parameter to beijing evn	6 months ago
📄 .gitignore	add blank	6 months ago
📄 pom.xml	init the code base,is not for use	6 months ago

大数据开放平台的灵魂



Maple-SDK

SDK For Integration Technical



大数据开发包-Maple-SDK

- ④ 数据建模 (DataSource)
- ④ Avro-Mapreduce编程库
- ④ Flume-ng扩展组件 (Flume-ng-ext)
- ④ 统计分析 (Maple-Report)
- ④ 分布式索引 (Maple-Index)

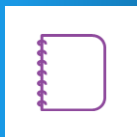


Data Source

大数据建模 (DataSourceec)

适用于大数据的动态、自动建模系统
实现数据导向理念的基础

用大数据的眼光看数据-DataSource



Partition

基本属性



Schema

文本格式

Avro格式

数据格式

列存储格式

数据库文件

HDFS

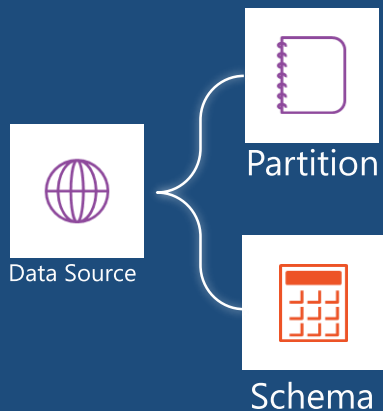
LocalFs

存储位置

DataBase

Memory

用大数据的眼光看数据-Data Source



- 一个字“多”
- 常用Partition策略
 - Hash Partition
 - 日期Partition
- 动态特性
 - 随Partition不同会变化
 - 不同Partition的Schema支持合并成新的Schema
- Schema的属性（适应并描述数据的变化特性）
 - 字段名
 - 字段稀疏性
 - 字段类型分布

围绕DataSource建立的数据导向API



HiveQL On Source



Spark Load Source



SharkQL On Source



Impala On Source



Pig On Source



实现融合的关键技术-Avro

Introduction

Apache Avro™ is a data serialization system.

Avro provides:

- Rich data structures.
- A compact, fast, binary data format.
- A container file, to store persistent data.
- Remote procedure call (RPC).
- Simple integration with dynamic languages.

Code generation is not required to read or write data files nor to use or implement RPC protocols. Code generation as an optional optimization, only worth in



开发
者

Thrift & Protobuf已经很成熟了，为
什么选择Avro？

Avro开发中代码生成是可选的，Avro
支持通用数据读取，更适应大数据变
化的特性。

有实践
的程序
媛



Avro在讯飞大数据开放平台的应用



Avro-Mapreduce
任务开发

- 高性能的数据序列化
- 简化的面向对象、富于设计的Mapreduce
- 支持Generic、Specific、Reflect (限于Java语言)



数据存储

- 支持通用数据读取
- 支持多种语言
- 内置多种压缩算法支持
- 与文本相比节省10倍存储空间
- 更高的读取性能



数据收集

- 多语言支持
- 与Flume-ng融合实现结构化日志收集
- 精简的数据格式，更高的数据传输速度



分布式结构化日志收集系统

- 部署节点超过1000个
- 每天收集千亿数据
- 用Avro封装了FlumeEvent，实现了结构化日志收集
 - 支持Log自定义结构体
 - 支持Log Array、Map等数据类型
- 得益于Avro，传输数据更精简，速度更快
- Flume-ng提供SDK，支持业务类功能扩展

围绕Flume-ng的优化

- ④ 以AvroFile为缓存的FileChannelPlus，极大的提升速度&稳定性
- ④ 支持Stable的改进版HDFS-Sink
- ④ 分布式节点监控&智能配置管理服务，弥补Flume-ng配置管理复杂的问题
- ④ 支持多语言的Loglib

结构化日志

多点监控





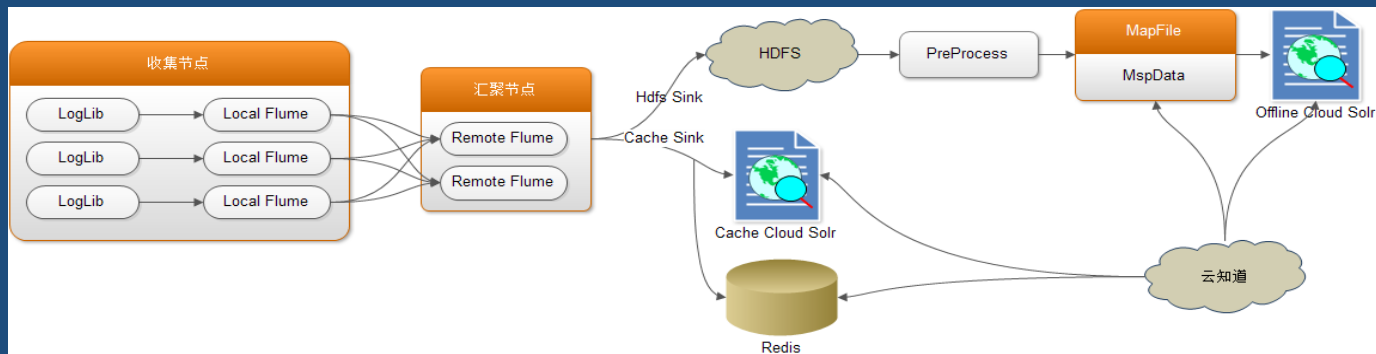
Avro-Mapreduce



实时日志检索系统

云知道数据流程图

日志收集与日志检索融为一体



云知道



千亿级别

- 目前每天日志索引记录15亿+
- 支持检索几个月数据，索引



即用即搜





Data
Source



Avro-Mapreduce



Sunflower语音云统
计分析系统



开放统计

专业免费的移动APP数据统计分析

<http://www.xfyun.cn/services/analysis/mobileapp>

讯飞开放平台统计分析

七大类，50多个小类统计分析功能，综合指标上千个

基于Hive的实现，分解后的Sql语句有上千条，运行太慢了

日2亿次PV，在语音重度服务下，日志量进千亿条



讯飞开放平台

基于Pig的统计分析脚本，也有好几百行，执行速度也很慢

优化的方向

对于同一份数据不同维度和指标的统计分析能否一次完成？

小时报表的计算结构能否被日报表利用，以此类推

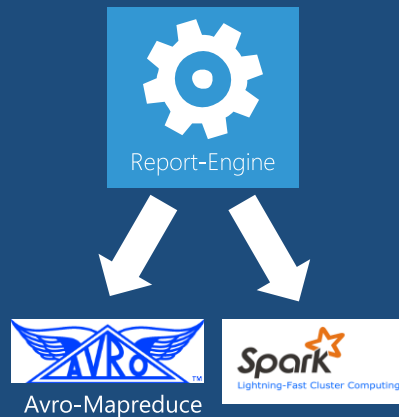
基于以上想法和对分布式计算原理的理解，我们开发了全新的统计分析解决方案
Maple-Report

统计分析解决方案 Maple-Report



Maple-
Report

- 体现数据导向理念
- 报表定义与计算引擎分离
- 同数据源的多维度、多指标一次计算完成
- 小时、日、周。。。数据依次复用



承载公司级大数据战略



Maple

数据汇聚



讯飞开放平台



讯飞输入法
说话就变文字



灵犀语音助手



酷音铃声

最后向那些以Doug Cutting为代表，依然耕耘在技术前线，勤于Coding的前辈致敬，是他们带给我们实实在在的大数据技术！

Q&A



讯飞开放平台

<http://www.xfyun.cn/>