五年总结：过往记忆大数据公众号原创精选

今年是我创建这个微信公众号的第五年，五年来，收获了6.8万粉丝。这个数字，在自媒体圈子，属于十八线小规模的那种，但是在纯技术圈，还是不错的成绩，我很欣慰。

我花在这个号上面的时间挺多的。我平时下班比较晚，一般下班到家了，老婆带着孩子已经安睡了，我便轻手轻脚的拿出电脑，带上耳机，开始我一天的知识盘点。

我写文，一方面是对自身技术知识的记录和积累，另一方面也是希望得到一些正反馈。一篇文章写出来，阅读量很多，动力当然就更大了。

对于本号，我坚持初心，只分享大数据技术文章，不掺杂别的内容进来。当然，广告我还是会接的，我不是富二代，上有老下有小，对金钱我有强烈的渴望。让我不图任何回报的熬夜写文章，而且坚持下来，这对我是很不公平的。大家如果不想看广告，跳过就行了。

下面我整理一下我这几年来自己写的原创文章，简略的分了个类（Spark、Hadoop、HBase、Flink、Kafka、CarbonData、Hive、ElasticSearch、大数据架构、分布式原理等），方便有需要的朋友收录。

如果想及时了解Spark、Hadoop或者HBase相关的文章，欢迎关注微信公众号：iteblog_hadoop

Spark篇

1、Spark & Alluxio在网易严选架构演进中的实践和探索
2、Apache Spark 中内存存储演进
3、深入理解 Spark Delta Lake 的诞生及其工作原理
4、Spark-SQL 在字节跳动的应用实践
5、一条 SQL 在 Apache Spark 之旅（上）
6、一条 SQL 在 Apache Spark 之旅（中）
7、一条 SQL 在 Apache Spark 之旅（下）
8、深入理解 Spark SQL 查询引擎

9、Airbnb 是如何通过 balanced Kafka reader 来扩展 Spark streaming 实时流处理能力的
10、Koalas: 让 pandas 开发者轻松过渡到 Apache Spark
11、.NET for Apache Spark 预览版正式发布
12、重磅｜Apache Spark 社区期待的 Delta Lake 开源了
13、Apache Spark 2.4 回顾以及 3.0 展望
14、SHC：使用 Spark SQL 高效地读写 HBase
15、从MPP数仓迁移至Spark：案例与最佳实践分享
16、Apache Spark 未来：Spark 3.0 预览
17、Apache Spark 3.0 将内置支持 GPU 调度，文末有福利
18、HBase 中加盐之后的表如何读取：Spark 篇
19、eBay：将60PB的MPP DBMS迁移至Spark的经验
20、Apache Spark 2.4 内置的 Avro 数据源实战
21、Apache Spark Shuffle I/O 在 Facebook 的优化
22、干货｜Spark 2.4 高阶函数介绍
23、Apache Spark 2.4 中解决复杂数据类型的内置函数和高阶函数介绍
24、SparkRDMA：使用RDMA技术提升Spark的Shuffle性能
25、MapReduce作业大规模迁移Apache Spark在百度的实践
26、Apache Spark 2.4 正式发布，重要功能详细介绍
27、即将发布的 Apache Spark 2.4 都有哪些新功能
28、Spark+AI Summit Europe 2018 PPT下载[共95个]
29、Spark Summit North America 201806 全部PPT下载[共147个]
30、Spark 从 Kafka 读数并发问题
31、Spark Streaming 反压（Back Pressure）机制介绍
32、Apache Spark 统一内存管理模型详解
33、干货｜Apache Spark 2.0 作业优化技巧
34、Apache Spark 2.3 重要特性介绍
35、Waterdrop：构建在Spark之上的简单高效数据处理系统
36、如何在 Hadoop 2.2.0 环境下使用 Spark 2.2.x
37、Spark作业如何在无管理权限的集群部署Python或JDK
38、Apache Spark 黑名单(Blacklist)机制介绍
39、Spark Summit 2017 Europe全部PPT及视频下载[共69个]
40、干货｜Apache Spark三大API：RDD、DataFrame和Dataset，我该如何选择
41、干货｜Apache Spark最佳实践
42、MMLSpark：微软开源的用于Spark的深度学习库
43、Apache Spark常见的三大误解
44、如何在Spark、MapReduce和Flink程序里面指定JAVA_HOME
45、Apache Spark 2.2.0新特性详细介绍
46、持续了半年的开发，Apache Spark 2.2.0今天正式发布
47、Spark Summit 2017全部PPT下载[共143个]
48、2017年Apache Spark两大发展方向：深度学习和提升实时流性能
49、Apache Spark常见的三大误解
50、如何优雅地终止正在运行的Spark Streaming程序
51、[资料]Spark Summit East 2017高清视频和PPT
52、Apache Spark 2.1.0正式发布，Structured Streaming有重大突破
53、Apache Spark又打破全球数据排序大赛记录
54、Spark Summit 2016 Europe全部PPT下载[共75个]

55、[Matei Zaharia]使用Apache Spark 2.0简化大数据应用程序开发
56、Apache Spark 2.0.1稳定版正式发布，可以考虑在线上使用啦
57、Hadoop&Spark解决二次排序问题(Hadoop篇)
58、Apache Spark 2.0.0正式发布及其更新介绍
59、Spark 2.0介绍：Spark SQL中的Time Window使用
60、Spark 2.0介绍：Catalog API介绍和使用
61、上海第九次Spark Meetup资料分享
62、Spark Summit 2016 San Francisco PPT免费下载[共95个]
63、杭州第四次Spark Meetup资料分享
64、Spark 2.0技术预览版正式发布下载
65、Spark 2.0介绍：Dataset介绍和使用
66、Spark 2.0介绍：SparkSession创建和使用相关API
67、Spark 2.0技术预览：更容易、更快速、更智能
68、Spark读取数据库(Mysql)的四种方式讲解
69、自定义Spark Streaming接收器(Receivers)
70、Spark Streaming和Kafka整合是如何保证数据零丢失
71、Apache Spark DataFrames入门指南：操作DataFrame
72、Spark读取数据库(Mysql)的四种方式讲解
73、Spark Checkpoint写操作代码分析
74、Spark社区可能放弃Spark 1.7而直接发布Spark 2.x
75、Spark分区器HashPartitioner和RangePartitioner代码详解
76、怎么在Idea IDE里面打开Spark源码而不报错
77、通过spark-redshift工具包读取Redshift上的表
78、Apache Hivemall:可运行在Apache Hive, Spark 和 Pig 上的可扩展机器学习库
79、Hadoop&Spark解决二次排序问题(Hadoop篇)
80、如何在Spark、MapReduce和Flink程序里面指定JAVA_HOME

Hadoop篇

1、Apache Hadoop 的 HDFS federation 前世今生
2、Hadoop 气数已尽？
3、如何从根源上解决 HDFS 小文件问题
4、在shell中如何判断HDFS中的文件目录是否存在
5、HDFS 块和 Input Splits 的区别与联系
6、Apache Hadoop 3.1.0 正式发布，原生支持GPU和FPGA
7、HDFS 副本存放磁盘选择策略详解
8、四种常见的MapReduce设计模式
9、三种恢复 HDFS 上删除文件的方法
10、Apache Hadoop 3.0.0 GA版正式发布，可以部署到线上
11、Apache Hadoop 3.0.0-beta1 正式发布，2017-11-01发布GA版即可在线上使用
12、Hadoop 3.0磁盘均衡器(diskbalancer)新功能及使用介绍
13、Hadoop集群字符集编码不一致导致Reduce重复记录问题排查
14、如何在Spark、MapReduce和Flink程序里面指定JAVA_HOME
15、NodeManager节点自身健康状态检测机制
16、NodeManager 生命周期介绍

17、使用CombineFileInputFormat来优化Hadoop小文件
18、MapReduce作业Uber模式介绍
19、Hadoop NameNode元数据相关文件目录解析
20、Apache Hadoop 2.8.0正式发布
21、Hadoop 3.0磁盘均衡器(diskbalancer)新功能及使用介绍
22、Apache Hadoop 3.0.0-alpha1正式发布及其更新介绍
23、Hadoop&Spark解决二次排序问题(Hadoop篇)
24、在shell中如何判断HDFS中的文件目录是否存在
25、设置Hadoop用户以便访问任何HDFS文件

HBase篇

1、HBase 中加盐之后的表如何读取：Spark 篇
2、HBase 中加盐之后的表如何读取：协处理器篇
3、HBase 协处理器入门及实战
4、HBase 入门之数据刷写(Memstore Flush)详细说明
5、为什么不建议在 HBase 中使用过多的列族
6、为了让你更全面的了解Apache HBase，我们做了这本专刊
7、HBase 性能和正确性调优
8、HBase Rowkey 设计指南
9、OpenTSDB 之 HBase的数据模型
10、HBase 多租户隔离技术：RegionServer Group 介绍及实战
11、OpenTSDB 底层 HBase 的 Rowkey 是如何设计的
12、HBase基本知识介绍及典型案例分析
13、Apache HBase中等对象存储MOB压缩分区策略介绍

Flink 篇

1、Flink Forward 201904 PPT资料下载
2、Flink Forward 201812 PPT资料下载
3、Flink Forward 201809PPT资料下载
4、Apache Flink状态管理和容错机制介绍
5、Flink在唯品会的实践
6、基于 Flink 的实时特征平台在携程的应用
7、Apache Flink 1.6.0 正式发布，涵盖多项重要更新
8、Apache Flink 1.5.0 正式发布，多项重要更新
9、四种优化 Apache Flink 应用程序的方法
10、Tumbling Windows vs Sliding Windows区别与联系
11、Flink Forward 201709所有PPT资料下载
12、如何在Spark、MapReduce和Flink程序里面指定JAVA_HOME
13、Apache Flink 1.3.0正式发布及其新功能介绍
14、[干货]Flink Forward 201704所有PPT资料下载
15、Flink四种选择Key的方法
16、Flink Table和SQL API：为统一批处理和流处理而设计
17、Apache Flink 1.2.0正式发布及其功能介绍

18、Flink可查询状态Queryable State:替换你的数据库
19、Apache Flink 1.1.0和1.1.1发布，支持SQL
20、Flink快速上手之Java API使用
21、Flink快速上手(QuickStart)
22、Flink是如何与YARN进行交互的
23、Tumbling Windows vs Sliding Windows区别与联系
24、[干货]Apache Flink 1.2.0新功能概述
25、[PPT干货]Apache Flink的未来
26、Apache Flink 1.1.3正式发布

Kafka篇

1、Kafka 是如何保证数据可靠性和一致性
2、图文了解 Kafka 的副本复制机制
3、Kafka创建Topic时如何将分区放置到不同的Broker中
4、重磅消息：Kafka 团队修改 KSQL 开源许可证，禁止其作为 SaaS 产品来提供
5、Kafka分区分配策略(Partition Assignment Strategy)
6、Apache Kafka 2.0.0 正式发布，多项重要功能更新
7、如何为Kafka集群选择合适的Topics/Partitions数量
8、图解Apache Kafka消息偏移量的演变(0.7.x~0.10.x)
9、Apache Kafka消息格式的演变(0.7.x~0.10.x)
10、Kafka创建Topic时如何将分区放置到不同的Broker中
11、Kafka分区分配策略(Partition Assignment Strategy)
12、Key为nulll时Kafka如何选择分区(Partition)
13、Kafka客户端是如何找到 leader 分区的
14、北京第三次Kafka Meetup活动PPT资料分享
15、Kafka实战：七步将RDBMS中的数据实时传输到Hadoop
16、Kafka Producer是如何动态感知Topic分区数变化
17、Kafka日志删除源码分析
18、Kafka集群Leader均衡(Balancing leadership)
19、Apache Kafka 0.10.0.0稳定版发布及其新特性介绍

CarbonData 篇

1、翻译 | Apache CarbonData 最新版中文文档发布
2、Apache CarbonData 1.0.0发布及其新特性介绍
3、Apache CarbonData的Update/Delete功能设计实现
4、Apache CarbonData性能基准报告：查询性能秒杀Parquet
5、Apache CarbonData快速入门编程指南
6、CarbonData：华为开发并支持Hadoop的列式文件格式

Hive篇

1、Apache Hive 联邦查询（Query Federation）
2、如何在 Apache Hive 中解析 Json 数组

3、Apache Hivemall:可运行在Apache Hive, Spark 和 Pig 上的可扩展机器学习库

ElasticSearch 篇

1、Open Distro for Elasticsearch：AWS 自家版本的开源 ElasticSearch
2、Elasticsearch 6.3 发布，你们要的 SQL 功能来了
3、ElasticSearch内置也将支持SQL特性
4、ElasticSearch 6.0即将发布，新特性展望

大数据架构

1、Uber 大数据平台的演进（2014~2019）
2、Alluxio在携程大数据平台中的实践
3、[干货]大规模数据处理的演变(2003-2017)
4、盘点2018年晋升为Apache TLP的大数据相关项目
5、盘点2017年晋升为Apache TLP的大数据相关项目

分布式原理

1、分布式原理：一致性哈希算法简介
2、分布式原理：一文了解 Gossip 协议
3、分布式快照算法: Chandy-Lamport 算法
4、一篇文章搞清楚什么是分布式系统 CAP 定理
5、大数据开发者应该知道的分布式系统 CAP 理论

其他

1、Apache Arrow：一个跨平台的内存数据交换格式
2、Apache Arrow：内存列式的数据结构标准
3、八个基本的 Docker 容器管理命令
4、Apache Zeppelin使用入门指南：编程
5、SSDB：可用于替代Redis的高性能NoSQL数据库
6、Scala的Option monad和C#的null-conditional操作符比较
7、Apache Beam发布第一个稳定版，适用于企业的部署
8、Scala模式匹配泛型类型擦除问题
9、Rheem：可扩展且易于使用的跨平台大数据分析系统
10、Flume+Morphlines实现数据的实时ETL
11、下一代大数据处理平台Apache Beam成为Apache顶级项目
12、Flume-ng禁用自动加载配置文件功能