火龙果•整理
uml.org.cn

2014 阿里技术沙龙
享技术·聚朋友

阿里HBase在搜索领域的设计与实践

徐斌 / 雨田

阿里巴巴高级搜索研发工程师

火龙果•整理
uml.org.cn

Agenda

- Overview
- Improvements
- Maintenance
- Extensional Projects
- Q & A

火龙果•整理
uml.org.cn

Overview

火龙果•整理
uml.org.cn

Upgrade History

Architecture

火龙果•整理
uml.org.cn

火龙果•整理
uml.org.cn

Features of our scenario

- YARN, HDFS and HBase coexist
- Intensive random read
- Various client type (JavaAPI, MR Job, iStream, Thrift …)

火龙果•整理
uml.org.cn

HBase Improvements

火龙果•整理
uml.org.cn

Increment Coprocessor

An incremental trigger mechanism to transmit application's real-time messages synchronously.

火龙果•整理
uml.org.cn

Other Assistant Coprocessors

Name	Description
Compare	Compare a Put data with the corresponding storage data, add additional data(constant value) into the Put if it matches some conditions.
Trace	Compare a Put data with the corresponding storage data, remove columns of Put if they are the same with the ones in the storage, leaving only the changed columns.
Copy	Compare a column data of a Put with some constant value (or other column data in the same Put), copy values to other columns according to configurations.

火龙果•整理
uml.org.cn

Thrift Server

- ✓ API improvements
- ✓ Scanner auto-release
- ✓ Add metrics to Ganglia

火龙果•整理
uml.org.cn

Constant Family Size Region Split Policy

- Set a constant limit for each family, if not, use the region max size limit instead.
- Region split will be triggered if any family reaches its size limit.
- The split point is determined by the family who exceeds the most proportion of its size limit.

For example:
SizeOf(F1) = 5M, SizeOf(F2) = 15M, SizeOf(F3) = 10M
LimitOf(F1)= 10M, LimitOf(F2) = 14M, LimitOf(F3) = 8M

火龙果•整理
uml.org.cn

Local Region Scanner

Open and read(scan) an online region at local machine in 'READ ONLY' mode, so that we can debug or trace some region issues without changing the region server's code.

火龙果•整理
uml.org.cn

Compaction Rate Limit

Check compaction process periodically. If it spends less time than we expected to compact a certain amount of data (for example: 10M), we should force it to 'sleep' a little while to lower the total compaction rate speed. (**HBASE-8329**)

火龙果•整理
uml.org.cn

Online Reload Configurations

MemStoreFlush, Split, HLog, Compaction and more…

火龙果•整理
uml.org.cn

Web-based Query Tool

1) Entry on Master's status page
2) Query by ROWKEY or URL
3) Support multi versions
4) Filter by column family

Table: pagebase

Master, Local logs, Thread Dump, Log Level

Query By Rowkey

Row Key : http://www.amazon.cn/ Query

Reversed URL URL

Max Versions : 1

Families : algorithm history image inlinks outlinks page_info parser raw selection trend uri

Family Name	Qualifier Name	Timestamp	Value
uri	db_time	1296087843883 Thu Jan 27 08:24:03 CST 2011	1296087519
	dir_depth	1310224558727 Sat Jul 09 23:15:58 CST 2011	0
	domain	1398259975925 Wed Apr 23 21:32:55 CST 2014	amazon.cn
	follow_depth	1298348855700 Tue Feb 22 12:27:35 CST 2011	63
	follow_type	1380092524325 Wed Sep 25 15:02:04 CST 2013	2
	host	1398259975925 Wed Apr 23 21:32:55 CST 2014	www.amazon.cn
	hub_depth	1326351653971 Thu Jan 12 15:00:53 CST 2012	0
	ip	1398259301153 Wed Apr 23 21:21:41 CST 2014	-883879550
	is_new	1398259301153 Wed Apr 23 21:21:41 CST 2014	0
	is_selected	1398259367056 Wed Apr 23 21:22:47 CST 2014	1
	link_depth	1326351653971 Thu Jan 12 15:00:53 CST 2012	0

火龙果•整理
uml.org.cn

Cluster Maintenance

火龙果•整理
uml.org.cn

Rolling Upgrade

Why do we need this?

HBase upgrade is always a nightmare for all users and administrators. We should find some ways to do it without a whole cluster shutdown.

When can we use it?

Rolling upgrade means upgrade region servers one by one (or group by group), so any IPC protocol version difference between region servers or between region server and master will cause catastrophic failures.

How did we use it?

Select a group of 10 ~ 15 region servers. Move all regions out of these nodes concurrently, shut these servers down, switch to new version, restart region servers, move back all regions to original locations.

火龙果•整理
uml.org.cn

Cluster Availability Control

- To block all the clients requests when HBase cluster is shutting down.
- Add a 'cluster unavailable' znode on HBase zookeeper. When it exists, all HBase clients(Java API, YARN App, ThriftServer) will be blocked.
- Use master's coprocessor to control the znode.

火龙果•整理
uml.org.cn

HBase Monitor (1)

Collect cluster's read & write requests status, display QPS graph using OpenTSDB.

Graph | Stats | Logs | Version

From 2014/04/21-12:00:00 To (now) Autoreload WxH: 1200x400

Every: 15 seconds

request +

Metric: request

Tags cluster offline-cm6

x table etao_comment

x type read

Rate Rate Ctr Right Axis

Rate Ctr Max:

Rate Ctr Reset:

Aggregator: sum

Downsample avg 10m

Axes Key Style

	Y	Y2
Label		
Format		
Range	[0:]	[0:]
Log scale		

431 points retrieved, 312 points plotted in 56ms.

火龙果•整理
uml.org.cn

HBase Monitor (2)

- Use a ‘region probe’ (actually, a special defined Scan operation) to request a single row from each region on each region server.
- Collect the response time, tag them with NORMAL or TIMEOUT.
- Draw a graph based on the proportion of NORMAL regions in all regions of a cluster using OpenTSDB.
- Send out warning messages(in form of Email, AliWangWang, SMS or VoiceCall) if the proportion of TIMEOUT regions of a single region server (or a single HTable) reaches a certain limit.

2014 阿里技术沙龙

火龙果•整理
uml.org.cn

Request Profiler

Real-time profiler for any slow request. Filtered by processing time or response size.

Call Tracker Current State 清屏

Enabled	ProcessingTime Threshold(ms)	ResponseSize Threshold(byte)	关键字过滤
true	20000	10485760	null
Close	20000 重设	重设	重设

process_time=20486 response_size=9576076 call_size=41 next(-2489589150756090708, 128), rpc version=1, client version=29, methodsFingerPrint=-1412189712 from 10.246.10.72:35718 ProfilingData:data_block_miss_cnt:1885, bloom_block_hit_cnt:68, index_block_hit_cnt:3155, index_block_miss_cnt:2, server_process_time.ms:20486, response_size.byte:9576076, server_queue_time.ms:0, total_block_read_time.ms:1876, HFILE_NAME:{115c0a4c6f75495c887524a6ccb8e0d2[DATA Block offset=56794/compressedSize=56943/decompressedSize=1013689],115c0a4c6f75495c887524a6ccb8e0d2[DATA Block offset=113737/compressedSize=56948/decompressedSize=1013635],115c0a4c6f75495c887524a6ccb8e0d2[DATA Block offset=170685/compressedSize=56957/decompressedSize=1013587],115c0a4c6f75495c887524a6ccb8e0d2[DATA Block offset=227642/compressedSize=56888/decompressedSize=1013877],115c0a4c6f75495c887524a6ccb8e0d2[DATA Block offset=284530/compressedSize=56717/decompressedSize=1013486],115c0a4c6f75495c887524a6ccb8e0d2[DATA Block offset=341247/compressedSize=56903/decompressedSize=1013719],115c0a4c6f75495c887524a6ccb8e0d2[DATA Block offset=398150/compressedSize=56726/decompressedSize=1013648],115c0a4c6f75495c887524a6ccb8e0d2[DATA Block offset=454876/compressedSize=57125/decompressedSize=1013286],115c0a4c6f75495c887524a6ccb8e0d2[DATA Block offset=512001/compressedSize=56671/decompressedSize=1013611],115c0a4c6f75495c887524a6ccb8e0d2[DATA Block offset=568672/compressedSize=56960/decompressedSize=1013732],115c0a4c6f75495c887524a6ccb8e0d2[DATA Block offset=625632/compressedSize=56989/decompressedSize=1013867],115c0a4c6f75495c887524a6ccb8e0d2[DATA Block offset=682621/compressedSize=56828/decompressedSize=1013573],115c0a4c6f75495c887524a6ccb8e0d2[DATA Block offset=739449/compressedSize=56751/decompressedSize=1013715],115c0a4c6f75495c887524a6ccb8e0d2[DATA Block

火龙果•整理
uml.org.cn

Region Server Read Rate Limit

- Launch region-level read rate control if the total size of the past 10s read operations reaches RegionServer's read limit(100M/s).
- Throw IOExceptions when a certain region reaches its own limit.
- Cancel the region-level control if all regions obey their limits in the past 10s.

火龙果•整理
uml.org.cn

Extensional Projects

火龙果•整理
uml.org.cn

HQueue (1)

What's HQueue?

It's a distributed and persistent message-oriented middleware based on HBase.

Features

- HQueue is based on HTable, so it also supports auto-failover, multi-partition(HTable region) and load balance.
- It's a persistent storage system. (HBase HLog & HDFS Append)
- High performance in both read and write.
- Message is classified by 'Topic'. (HBase column qualifier)
- Support TTL mechanism. (HBase's KeyValue level TTL)
- As a lightweight wrapper of HTable, HQueue can upgrade with HBase cluster seamlessly.
- Allows map/reduce jobs to assign tasks based on locality.
- Support subscribe mechanism.

火龙果•整理
uml.org.cn

HQueue (2)

火龙果•整理
uml.org.cn

OpenTSDB

TSD Time Series Database

What's OpenTSDB?

OpenTSDB is an open-source, distributed time series database designed to monitor large clusters of commodity machines at an unprecedented level of granularity. OpenTSDB allows operation teams to keep track of all the metrics exposed by operating systems, applications and network equipment, and makes the data easily accessible.

Performance

Thanks to HBase's scalability, OpenTSDB allows you to collect thousands of metrics from tens of thousands of hosts and applications, at a high rate (every few seconds). OpenTSDB will never delete or downsample data and can easily store hundreds of billions of data points.

Where do we use it?

Etao Search Dump Stats, Web Crawler Stats, HBase Monitor and more …

火龙果•整理
uml.org.cn

Phoenix

Phoenix
"We put the SQL back in NoSQL"

What's Phoenix?

Apache Phoenix is a SQL skin over HBase delivered as a client-embedded JDBC driver targeting low latency queries over HBase data.

Principle

Apache Phoenix takes your SQL query, compiles it into a series of HBase scans, and orchestrates the running of those scans to produce regular JDBC result sets.

Performance

Direct use of the HBase API, along with coprocessors and custom filters, results in performance on the order of milliseconds for small queries, or seconds for tens of millions of rows.

火龙果•整理
uml.org.cn