HBase Today

November 19th, 2010

Michael Stack
StumbleUpon

Overview

- HBase Today
- Notable Deploys
- Notable Add-ons

Me

- ~~Chair~~ Janitor for HBase Apache Project
- Member of Hadoop PMC
- stack@apache.org

HBasics

- Apache TLP
- Distributed, column-oriented store
- Based on Google BigTable paper [2006]
- Built on Hadoop Core, HDFS & ZooKeeper

HBasics

- HBase is all about……..

- Near linear as you add machines
  - ‘Sharding’/‘Partitioning’ is built in from get-go
    - Autosharding
  - BigTable known to scale into the thousands of nodes
  - HBase Project Goal
    - Billions of rows X millions of columns on clusters of "commodity" Hardware

HBase Yesterday

- Current Stable release 0.20.6

HBase Today

- HBase 0.90
  - Next major version after HBase 0.20
  - Break with Hadoop versioning
  - Closer to 1.0!
- Prefaced by three 0.89.x ‘developer releases’
- RC0 posted Monday, 15th November
  - 930 issues closed since 0.20

HBase 0.90

- Lots of work tightening up ACID guarantees
  - No read of half-written row
- Durability
  - Requires HDFS with sync support
    - Flush edits through write pipeline before returning to client
    - branch-0.20-append OR CDH3 (b2 or b3)
  - Tunable
    - Group commit
    - Deferred log flush
      - Flush every N ms

HBase 0.90

- Testability refactoring
  - Mockable Server/Services Interfaces
  - 60 new test classes
  - 13k lines of new test code
  - Miniclusters of ZK ensembles, HDFS, multiple masters and regionservers all in one JVM

HBase 0.90

- Rewrite of Master process
  - All cluster state moved out to zookeeper
  - Master failover
  - Concurrent failure of Master + Regionservers

Replication

- Edits from one HBase cluster to another
  - Geographically distributed or not
- Why?
  - Earthquake ate my HBase!
  - Synchronization
    - “Live” to “science” cluster
      - And back again
- WAL shipping
  - State of replication kept up in ZooKeeper

Bulk Load

- Load MapReduce output up into HBase
  - 10-100x improvement over API
- MapReduce job writes HFileOutputFormat
  - TotalOrderPartitioner
    - Or custom partitioner
  - *$ bin/hbase completebulkload /output-path TABLE*

HBase 0.90

- Mavenized
- HBCK
  - Some repair facility
- Performance
  - Compaction algorithms
  - Blooms done right
  - Multi-Put, Multi-Get, Multi-Delete
  - Efficient reads in rows of millions of columns/versions
    - Seek-forward

HBase 0.91

- Next month, new “developer release” series

HBase 0.92

- Coprocessors
  - Arbitrary code that runs at each Region
- Security
  - Pull Hadoop security up into HBase
- Multi-master replication
- Better balancing
  - read/write load
  - data locality
- Online schema changes
  - Schema in zookeeper

Security

- Per Column Family
- ACL
  - Read/Write/Execute/Create/Admin
  - SuperUser
  - Multi-tenant
- Kerberos Client Auth
  - Hadoop RPC via SASL
  - Optional Encryption
- Support in Shell
  - Config Security Schema

Coprocessors

- Framework for building distributed services
- Two styles
  - Pre/post hooks on all Region operations
    - get, put, scan, delete
    - open, close, compact
  - Arbitrary operation on Region
    - Dynamic RPC, define own client/server protocol
- CPs can be chained
- Classes loaded from HDFS
  - Loaded on Region open
  - Dynamically load to running cluster

Coprocessors Examples

- Security
- Aggregate functions
  - sum, avg, etc.
- Secondary indexing
- Region indexing
- Online MapReduce
  - Runs concurrently on all cluster nodes

Notable Deploys

StumbleUpon

- *"A discovery engine that finds the best of the web, recommended for each unique user"*
- *12M users*
  - *Growing*
- 500M "stumbles" a month
- Stumblers spend on avg. 7 hours a month "stumbling"
- Big driver of traffic to other sites
  - More than digg, del.icio.us, slashdot, etc. combined

StumbleUpon

- Multiple clusters
  - Multiple datacenters
    - Online
    - Offline processing
- Replication
  - backup
  - Recommendations team has near-live data
- Counters
  - real-time analytics
    - 10s/100s of thousands of increments a second
- PHP to HBase via thrift

Mozilla

- Mozilla Socorro
  - –Store and process Firefox crash reports
- Receive about 3.5 million reports a day
  - –~450GB/day
  - –Growing
    - Cluster filling at 3x predicted rate
    - New, bigger cluster!
- Old system was only able to process 15%
- Reads and Writes directly to HBase
  - –Python clients via thrift

yfrog

- Service run by imageshack
  - –lets you“share your photos and videos on twitter”
- 60 servers running datanodes
  - –30 RegionServers
  - –30 TaskTrackers for bulk loading
  - –Intel(R) Core(TM)2 Duo CPU E7300 @ 2.66GHz
    - 6 disks
- 80TB of images => PB
  - –15k regions
  - –HTTPD w/ Varnish and inline ImageMagick
    - Backed by HBase

Miscellaneous

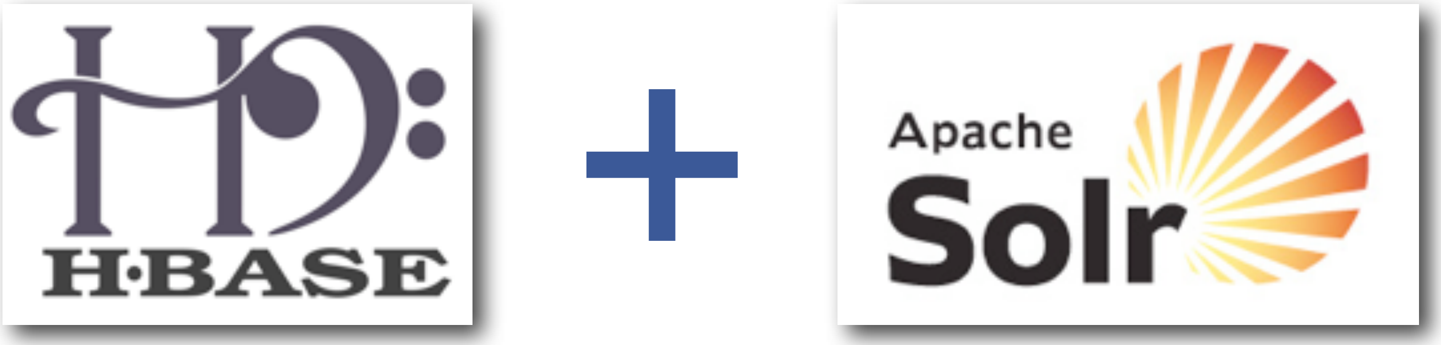
- twitter
  - Backend Analytics
- facebook
  - Talk tomorrow @ 15:10
- YAHOO!
  - COKE Analytics team
  - Others @ Y! => 650 node HBase cluster
- Adobe
  - Analytics Platform

Notable Additions

- OpenTSDB
- Lily Project: Content Repository
- MySQL to HBase replication

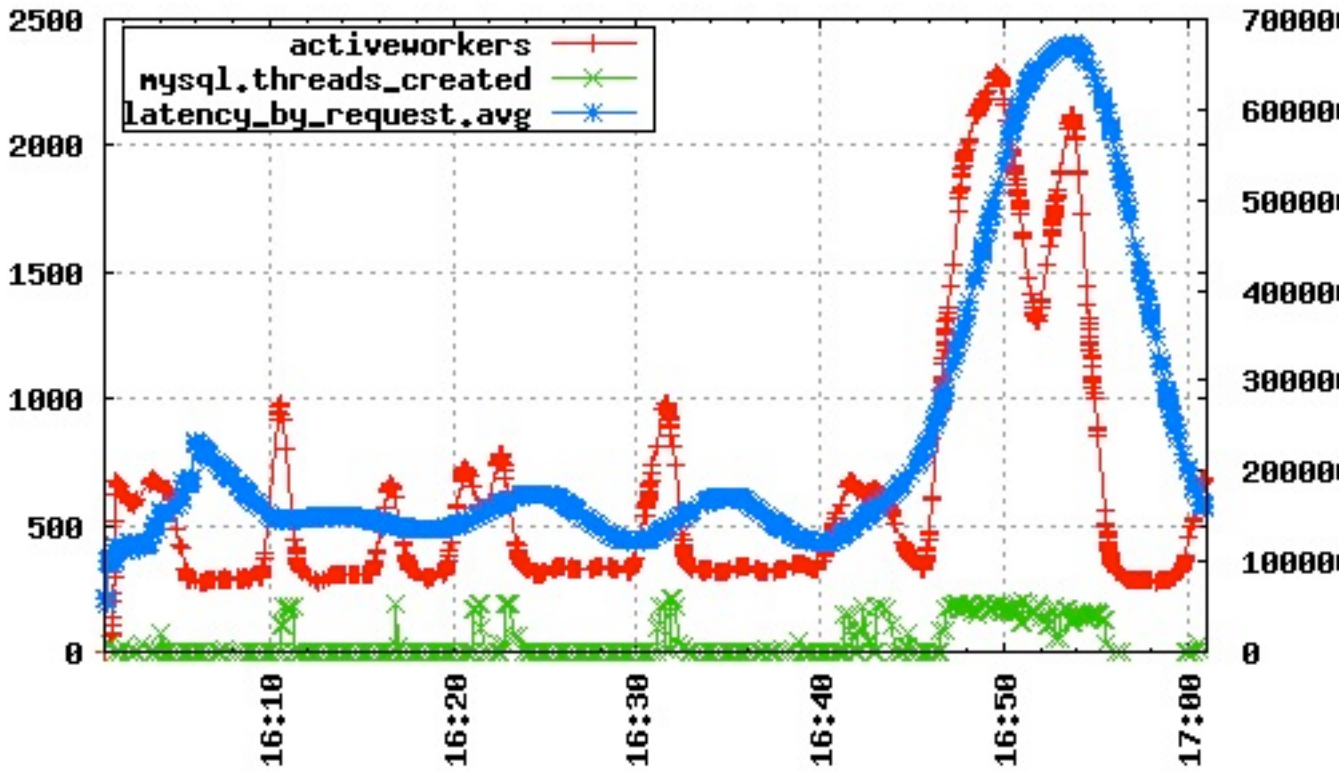
lily

- Scalable Content Repository
- Open-source (ASL)
- Data stored and archived in HBase
  - HBase Indexing Library for secondary indexes
    - Incremental or MR batch build
  - HBase RowLog for SOLR WAL and reliable queueing



OpenTSDB.net

- Distributed, scalable time-series DB
  - –Ganglia on Steroids
  - –Built on HBase
- Shared nothing daemons
  - –chipmunk to ride over outages
  - –1B datapoints a week @ SU
  - –AsyncHBase Client
  - –Twisted Deferred Pattern
- Graphs in real-time
  - –aggregates
  - –averages



MySQL Replication

- MySQL to HBase replication
  - RBL
  - Based on Tungsten Replicator
- For
  - MySQL backup
  - Migration to HBase
  - Analytics in HBase
- OSS soon
  - Dec2010

Thanks

- hbase.org
- stack@apache.org