SERENDIO

Smart Meter Big Data Analytics
The Business Case for Hadoop

Prepared by Kregg Ray @ Serendio

October 2014

Executive Summary

MongoDB, a document data base, is not well suited for the time series data analytics at the crux of [client's] smart meter analytics business model. Hbase, a columnar database is. Here's why it makes sense to abandon the MongoDB benchmark, and some supporting arguments to instead move ahead on Hbase.

Background

Over the past 10 years [Client] has established a reputation as a pioneer and industry leader in utility fraud detection and loss prevention analytics. [Client] is experiencing rapid growth and increased demand for a broader range of analytics and more timely operational and customer intelligence as utilities quickly deploy new smart meters.

[Client's] current SQL platform has known limitations in terms of scalability and performance in the era of Big Data smart meter analytics. [Client] is experiencing competitive pressure (losing RFPs to vendors with Big Data support) to offer near real time operational and customer intelligence for enhanced demand response programs, regulatory compliance, AMI network analysis, new pricing models and service offerings, and the reduction of revenue loss from theft and fraud.

To remain competitive, [Client] must redesign the analytics platform to support ingestion of smart meter data at intervals of every 15 minutes, or less if possible, as opposed to the weekly or daily intervals data is being collected today. Additionally, a broad range of new analytics and operational reporting will be necessary to meet market demand.

Due to the lack of internal hands-on NoSQL skills and resources specialized in Big Data technology, [Client] has conducted its own internal (paper based) evaluation of Cassandra and has disqualified it based on [confidential]. Simultaneously, [client] has engaged an independent consultant in a different state to conduct a benchmark of MongoDB with a representative sample of 12 months of customer smart meter data (hundreds of columns and hundreds of millions of rows) executing a selected sample of core business analytical queries. Due to various circumstances, the MongoDB benchmark has stalled.

Client has engaged Serendio to get the platform selection project moving forward again and assist in the evaluation and selection of the ideal Big Data platform by; validating the benchmarking and research done to date by [client] and [external consultant], assisting in the definition and articulation of a clear operational business strategy for how best to leverage, deploy and manage Big Data technology for smart meter analytics, benchmark Hbase with the same test harness used for MongoDB, and make recommendations and offer a proposal to design, develop, deploy and manage the smart meter analytics platform under a managed services engagement.

NoSQL Landscape

Key-Value Store (Redis, Riak)
Document Store (Mongo, Couchbase)
Column Store (Cassandra, Hbase)
Graph Store (Neo4j, Node.js)
Examples, When to use which and why

Focus on Document Store and Columnar

MongoDB and Couchbase

Use cases: user profiles, product catalogs, geospatial, financial products (deep nests), social media, digital content, gaming, metadata, events, bills and invoices

Hbase and Cassandra

Use cases: structured, semi-structured, unstructured data, full table scans, read intensive operations, time series interval data, geospatial data

Columnar Database Analysis

CAP Theorem – Concepts and Misconceptions
- Partition Tolerance
- Consistency vs. Eventual Consistency
- Availability

SQL Friendliness
- CQL in the Shell
- Co-Processors, stored procedures, triggers
- Phoenix and Impala - JDBC Drivers
- Splice Machine (Hadoop on ACID)

Complexity (One man's complexity is another's modularity)
- Hadoop Ecosystem
  - MapReduce (Batch vs. Near Real Time - Hadoop 2.0)
  - Storm, Spark, Shark
  - Oozie, Flume, Sqoop
  - Hive, Pig

SPoF and Hot Spots
- NameNode vs. Peer-to-Peer
- Gossip, YARN and Zookeeper
- Hortonworks and Cloudera HA solutions
- Hotspots and proper key construction

Market Landscape

Who's backing Hadoop and why?
- Sponsors, Committers, Clients
  - Google, IBM, Microsoft, HP, LinkedIn, Twitter, Facebook, Intel, **OpenTSDB**
  - Hortonworks, Cloudera

Who's backing Cassandra and why?
- Sponsors, Committers, Clients
  - Netflix, Rackspace, Facebook, Twitter, Apple, eBay, ConstantContact
  - Datastax

Utility Industry and Big Data

Utilities Uncertain about Big Data Analytics

What Utilities are Investing in: Big Data Analytics

Benefits of Big Data to Utilities

How Utilities Profit from Big Data

Smart Meter Analytics Challenges for Utilities

Competitive Landscape

OPower (Hbase)

C3 (Hbase)

Pulse Energy (Cassandra)

AutoGrid (Hbase)

Serendio Recommendations

Abandon MongoDB benchmark

Benchmark current test harness on Hadoop production scale infrastructure

Benchmark current SQL deployment to Hadoop

Design and implement Competitive Smart Meter Analytics Platform MVP

Competitive Smart Meter Analytics Platform

Design and Architecture for both Operational and Business Intelligence

Data Sources

ETL

Storage and Persistence

Analytics

Visualization

MVP

Evolve and enhance the current POC to a Minimum Viable Product (MVP) to meet [client's] key business requirements.

Timeline: 4-6 months (this can be adjusted based on scope and deadlines)

Resources: 1 Architect/Project Manager, 3 FT Data Engineers

Managed Services

Manage the Big Data infrastructure; provide ongoing customization, enhancement, maintenance and support

Resources: 0.5 Project Manager, 1 FT Support Engineer/Sys Admin, 1 FT Data Engineer

References

NoSQL Landscape

NoSQL landscape: http://blogs.the451group.com/information_management/2011/04/15/nosql-newsql-and-beyond/

Exploring different types of NoSQL databases: http://www.3pillarglobal.com/insights/exploring-the-different-types-of-nosql-databases

When and why to use Hbase, Mongo, Cassandra
http://www.slideshare.net/EdurekaIN/no-sql-databases-35591065

Columnar vs. Document store databases:

Columnar databases for dummies: http://www.dummies.com/how-to/content/columnar-databases-in-a-big-data-environment.html

Document database compared to columnar database:
http://stackoverflow.com/questions/15294507/scenario-for-document-vs-columnar-dbst

Ten common tasks for MongoDB: http://www.infoworld.com/article/2612785/application-development/10-common-tasks-for-mongodb.html

MongoDB use cases: http://docs.mongodb.org/ecosystem/use-cases/

MongoDB sells "popularity" as a benefit: http://www.mongodb.com/leading-nosql-database

Mongo Hbase side by side: http://db-engines.com/en/system/HBase%3BMongoDB

Excellent LinkedIn discussion – why not to compare Mongo to Hbase:
https://www.linkedin.com/groups/HBASE-MONGODB-4531843.S.172759461

Hbase v. Cassandra

Hbase complexity argument neutralized by Datastax adding support for Hive, Pig and Storm, Spark, Shark: http://planetcassandra.org/getting-started-with-apache-spark-and-cassandra/?gclid=CMv7zuH5lcECFZSFfgodLSQAlw

Side by Hbase Cassandra technical comparison:
http://bigdatanoob.blogspot.com/2012/11/hbase-vs-cassandra.html

JavaWorld Review Hbase v Cassandra:
http://www.greentechmedia.com/articles/read/c3_smart_grids_biggest_big_data_contender

Debunking CAP "choose any two": http://stackoverflow.com/questions/15303343/relational-vs-columnar-and-document-databases-arent-they-one-in-the-same

Consistency vs. Eventual Consistency: http://stackoverflow.com/questions/12222469/why-opentsdb-chose-hbase-for-time-series-data-storage

Hadoop not built for transactions: http://www.slideshare.net/enissoz/hbase-high-availability-for-reads-with-time

Who's using Hadoop:

Hbase committers: http://www.slideshare.net/enissoz/hbase-high-availability-for-reads-with-time

IBM:
http://www.theregister.co.uk/Print/2013/04/03/ibm_puredata_hadoop_appliance_biginsights/

Microsoft: http://azure.microsoft.com/blog/2014/08/25/azure-hdinsight-makes-hbase-nosql-database-a-ga-feature/

Yahoo: https://developer.yahoo.com/blogs/ydn-blog/apache-hbase-yahoo-multi-tenancy-helm-again-203911418.html

Intel (invests in Cloudera): http://www.informationweek.com/big-data/hardware-architectures/intel-invests-in-cloudera-but-what-changes/d/d-id/1141573

eBay: http://www.slideshare.net/Hadoop_Summit/ma-june27-140pmroom212v2

LinkedIn: http://www.infoq.com/news/2010/08/linkedin-data-infrastructure

LinkedIn's solution to Big Data: Hadoop: http://www.zdnet.com/linkedins-answer-to-big-data-problems-pinot-7000034059/

HP: http://www.informationweek.com/big-data/software-platforms/hp-invests-$50m-in-hortonworks-hadoop-bet/d/d-id/1297542

Who's using Cassandra:
Yahoo Japan: http://planetcassandra.org/blog/yahoo-goes-woohoo-for-apache-cassandra-cassandra-wins-yahoo-japan-nosql-evaluation-with-lowest-latency-and-highest-scalability/

Netflix: http://www.datastax.com/oracle

Datastax secures $106M: http://thenewstack.io/armed-with-additional-106-million-datastax-to-keep-pushing-cassandra-to-enterprises/

Datastax secures $45M: https://gigaom.com/2013/07/23/nosql-startup-datastax-raises-45m-to-ride-cassandras-wave/

Columnar Database Analysis

InfoWorld Hbase Review: http://www.infoworld.com/article/2610709/database/review--hbase-is-massively-scalable----and-hugely-complex.html

Storm, Spark, Shark: http://planetcassandra.org/getting-started-with-apache-spark-and-cassandra/?gclid=CMv7zuH5lcECFZSFfgodLSQAlw

Storm and Spark contrasts: http://stackoverflow.com/questions/24119897/apache-spark-vs-apache-storm

Storm promoted to Top Level Project:
https://blogs.apache.org/foundation/entry/the_apache_software_foundation_announces64

Hbase HA argument (Cloudera and HW both have fixes):
http://www.slideshare.net/cloudera/120613-hadoopsummithbaseavailabilitybean-hsieh

Hortonworks Hbase HA: http://www.slideshare.net/enissoz/hbase-high-availability-for-reads-with-time

Good explanation of hotspots: http://stackoverflow.com/questions/15294507/scenario-for-document-vs-columnar-dbs

Hotspots in Hbase (Apache doc on how to avoid them): http://hbase.apache.org/book/rowkey.design.html

Knocks on Hbase – Hadoop v1 v v2: http://www.slideshare.net/EdurekaIN/edureka-hadoop2-architecturewebinar-34201277?related=1

Hadoop v1 vs v2: http://www.slideshare.net/EdurekaIN/edureka-hadoop2-architecturewebinar-34201277?related=1

Hbase SPoF myth: http://www.smartgridnews.com/artman/publish/Delivery_Asset_Management/Data-analytics-buying-guide-part-1-What-s-in-Big-Data-for-you-5253.html#.VDAMixZ_Tjk

Does Hbase scale? French POC: http://www.slideshare.net/Hadoop_Summit/proof-of-concent-with-hadoop

What the Market is saying

back and forth Cassandra and Hbase: http://www.informationweek.com/big-data/software-platforms/big-data-debate-will-hbase-dominate-nosql/d/d-id/1111048?

Why FB switched: http://www.quora.com/Why-did-Facebook-pick-HBase-instead-of-Cassandra-for-the-new-messaging-platform

Why FB switched: Hbase good for read write, Cassandra good for fast write (why FB switched): http://stackoverflow.com/questions/23422181/cassandra-good-for-write-and-less-read-hbase-random-read-write

Comparison of Mongo, Cassandra and Hbase, mentions Splice Machine and Microsoft recent developments: https://gigaom.com/2014/08/10/is-hbases-slow-and-steady-approach-winning-the-nosql-race/

Why OpenTSDB chose Hbase (well suited for time series data scans): http://stackoverflow.com/questions/12222469/why-opentsdb-chose-hbase-for-time-series-data-storage

Hbase and Cassandra co-exist – not mutually exclusive, so says Datastax: "we do have customers that use more than just Cassandra (C*). On our customers page you'll find examples like MarkedUp (all 3), eBay (C* and Hadoop), Datafiniti (C* and Solr), HealthCare Anytime (all 3), Constant Contact (C* and Hadoop), SimpleReach (C* and Hadoop), Boxever (C* and Hadoop), and Skillpages (all 3)."

http://www.slideshare.net/EdurekaIN/no-sql-databases-35591065
http://www.javaworld.com/article/2140805/big-data/big-data-showdown-cassandra-vs-hbase.html

Switched from Hbase to Mongo – Trakker: http://traackr.com/blog/2012/02/traackrs-migration-from-hbase-to-mongodb/

Challenges faced by Utilities

Low adoption rate of Hadoop in Utilities industry, mostly due to lack of available skills: http://www.smartgridnews.com/artman/publish/Business_Analytics/Survey-reveals-utilities-are-messing-up-analytics-6034.html#.VDRZmhZ_Tjk

Utilities Uncertain about Big Data Analytics: http://tdwi.org/Articles/2014/08/05/Utilities-Uncertain-about-Big-Data-Analytics.aspx?Page=2

How Utilities are profiting from Big Data: http://eandt.theiet.org/magazine/2014/01/data-on-demand.cfm

What Utilities are investing in – Analytics: http://www.slideshare.net/Hadoop_Summit/ma-june27-140pmroom212v2

Smart meter analytics challenges for utilities: http://www.utilityanalytics.com/resources/insights/analytics-post-smart-meter-world

Big Data benefits for utilities: http://www.smartgridnews.com/artman/publish/Delivery_Asset_Management/Data-analytics-buying-guide-part-1-What-s-in-Big-Data-for-you-5253.html#.VDAMixZ_Tjk

Big Data Insights for Utilities: http://www.intelligentutility.com/article/13/10/big-data-insights-smart-utilities

Competitive Landscape

OPower Deck: http://storage.pardot.com/17572/37408/Converting_Your_Smart_Grid_Data_into_Real_Customer_Value_20120228_2101_1.mp4

OPower on Hadoop: http://www.providencejournal.com/business/press-releases/20141001-pepperdata-enables-opower-to-rely-on-hadoop-for-real-time-big-data-analytics.ece

OPower and AutoGrid both using Hadoop: http://www.greentechmedia.com/articles/read/opower-takes-on-big-data-for-home-energy

AutoGrid raises $12.75M: http://finance.yahoo.com/news/autogrid-systems-raises-12-75-120000613.html

Pulse Energy: http://www.pulseenergy.com/company/contact-us/

C3: http://www.greentechmedia.com/articles/read/c3_smart_grids_biggest_big_data_contender

AWS capacity planning

http://www.smartgridnews.com/artman/publish/Delivery_Asset_Management/Data-analytics-buying-guide-part-1-What-s-in-Big-Data-for-you-5253.html#.VDAMixZ_Tjk