Big Data Platforms Galore: Lessons from Marketing Associates

Presented by Andy Frey
CIO
Marketing Associates and Magnify Analytic Solutions
Overview

- Who is MA?
- Galore....
- What's out there
- The Big Data Choices
  - How do I choose - Selection Considerations
  - Cost matters - it's never just about the technology
- Selecting your new Data Store
  - Open Source vs Supported Open Source
- Where do I run it? Public or Private Cloud, Hyperconverged or Hybrid
- People & What Else
- The winners are...
Marketing Associates... is a full-service, technology-enabled marketing and analytic services company headquartered in Detroit. Offices in Charlotte, NC, Wilmington, DE, Shanghai, PRC, and Cologne, Germany

was established in 1967 and for 50+ years has produced high-quality integrated marketing solutions for Fortune 500 companies across different industries

is ISO 27001 Certified with a commitment to continuous improvement and quality delivery

believes that the only marketing strategies and programs that drive real, measurable results are those that begin from a deep understanding of our customers’ business challenges and objectives

is flexible, responsive and proactive. We deliver world-class sales and marketing programs on time and on budget

is a trusted partner to our clients with a SSAE-18 data center/program HQ for hosting and maintaining client applications and data
Some of our Customers

Automotive
- Ford
- Advance Auto Parts
- Daimler
- Michelin
- Navistar
- Continental

Financial / Insurance Services
- Amerisure Insurance
- Citizens Financial Group, Inc.
- Alphera Financial Services
- Delaware Lottery
- Ford Credit

Healthcare
- GE Healthcare
- McKesson
- Saint Francis Healthcare

Industrial Products
- Kuka
- Rockwell Automation

Life Sciences
- DuPont
- Syngenta
- The Tapco Group
- Chemours

Consumer Products
- KitchenAid
- Whirlpool
- Comcast
- Arby’s
- Hush Puppies
The largest list of data stores ever....

Many Many Hadoop variations & many NoSQL and a lot of RDBM(s)

Hadoop solid but may no longer be the default big data solution

(R)DBMS vendors offer big data platforms/integration they still want your money
  - Microsoft, Oracle, IBM, HP, AWS, Azure offer have big data extensions
  - More mature security integration with HDFS and NoSQL data stores
  - Don’t forget PostgreSQL, MYSQL, other if you use Open Source

Cloud Offerings a good way to get started
  - You still need to do your homework and pick the right tools and platform

Many analytical solutions use both RDBMS and a HDFS or MongoDB like data store

NoSQL database technology MongoDB is proven. Excels when data is ever changing
So many Choices

What’s Out There
How Do I Choose?

Familiar names in the “magic quadrant”

- Data Management Solutions for Analytics

Where is the Apache Stuff? Spark? Storm? Buried in the Solutions from the commercial Distributions.
Do I need a Big Data solution? Eventually at least…

How else can you sell to credit worthy clients, optimize JIT manufacturing, reduce operational costs, predict sales outcomes, monitor – predict customer defection, etc…

Define your accomplishment or goal – you need to do something impactful

Define your Use Case – Operational, Analytical, Interactive (real time) – all

Before you decide – analyze your data Volume, Variety, Velocity, Veracity, Viscosity and some other V words…… know your data!

data use case should impact your platform choice

What you are good at now should influence your selection

- Consider your legacy vendor – SQLServer, Oracle, others integrate with big data platforms
- May make staff more comfortable and will add to skills not revolutionize skills
- If an RDBMS will satisfy your use case – its ok – that’s why they have been relevant for 30 years. Remember they scale vertically not horizontally – can add a “big data” store later
Apache Spark has momentum
  • Hadoop users frequently replacing MapReduce with Spark for Hi speed In Memory functions
  • Apache Spark is Versatile – operational and analytical, Visualization, Interactivity, ML, IoT

Apache Storm good for real time analytics

The hardest area to change in an application is in the data store/organization

Plan carefully. If building an application – make sure you need a Big Data data store? If analytical or analytical interactive/streaming than Yes Big Data is for you.

Technical Envy (Developers Remorse) – unless you made a gross mistake – stick with choice – there is always another Apache project right around the corner to make you feel you should have picked that instead.
How Do I Choose?

Where does my data store fit?

Getting one that does one of these very well is hard.

You can write data fast but can you read it fast?

Supported Distributions “subscriptions” are comparable to legacy data store licensing costs – and they cost even more over time.

Operational Ease takes a veteran staff.

#5 – How close can it come to the big data “bulls-eye”?

Getting a big data technology that provides two out of three can be challenging; finding one that supplies all three can be very hard.
Isn’t everything in this world almost always about money?
- Develop a cost model with Licensing, Hosting, and Staff
- DIY Open Source solutions? Do you have the staff to compile and integrate all the tools? Security, Monitoring, Availability, Bugs…
  - Ok if you have a great System Admin/Developer and an Analytical use case with no PII and no deadlines - do not see this as viable for most enterprises
- Commercial Open Source Distribution is a requirement if Enterprise operational
- Ensure your solution has a way to encrypt your data – especially if using (s)PII
- Tool should have reasonable way to handle Data Access (ACL)
Cloud vs. On Premise

Unless you have a sophisticated datacenter and staff best to start on a ready made cloud or Private Cloud (MA) offering if new to Big Data

On Premise or private cloud is still cheaper than public cloud last time I checked....

Know how you are getting your data to and from the cloud.

Moving data in and out of cloud costs $$

If on premise - Hyperconverged or Hybrid platform?

Hyper is easier to manage but may wait for new features

Hybrid harder to manage but can integrate new features fast

If on premise – or DIY cloud

Separate compute from storage!!!!

Allows you to scale your compute and storage independently

Better performance and more flexibility

Can replace the tool platform you use but leave the data alone
Do you have the staff to run big data?

Java, Python, R, and SCALA skills in demand

Need an Architect Doer (HW, OS, Network, Security, DB, Java)

Need a Sr. Developer Doer (Java, Python, Scala, SQL, R)

Need a Data Analyst (ok Scientist) (R, SAS, Python)

- Experienced resources difficult to find
- Hire technical resource (at least 1) with proven experience
- Train your staff on your solution - preferably platform foundation class - offsite.

Hire from universities like **WSU** for interns and new hires

It only takes about 3 months or less for meaningful work to start

Take a chance - a good find is more than worth it
The Big Data Choices

What else is relevant

• GPDR & China Data PI.. looming
  • Immediate concerns depending on your business and data locality
  • Almost impossible to handle in the Big Data stores
  • Very difficult to handle in RDBMS
  • May make you migrate your Cloud data

• Encrypting PII
  • Difficult to do with HDFS
  • Need to design from beginning

• Access Control and Security
  • Big Data Offerings have much more maturity and granularity to gain
  • Commercial subscriptions like Cloudera Hadoop help mitigate exploit exposure
  • All systems require frequent patches
The Big Data Choices

The Winners are

• Apache Spark best overall tool
• Best fit for versatile workloads
  • Machine Learning, IoT, Analytical,
• Streams fast
• Ready for IoT

Scala best Language

IoT will be the fastest growing use case

Separate compute from the data!!!!
Flashblade is best of breed

5th Annual Big Data & Business Analytics Symposium - March 22-23, 2018