Managing Successful Data Projects: Technology Selection and Team Building

Strata Data Conference, New York 201
Ted Malaska | @ted_malaska
Jonathan Seidman | @jseidman
About the presenters

Ted Malaska

- **Capital One**: Director of Enterprise Architecture (Starting Next Week)
- **Blizzard Ent**: Director of Engineering of Global Insights
- **Cloudera**: Principal Solution Architect
- **FINRA**: Lead Architect
- **Contributor**: Apache Spark, Hadoop, Hive, Sqoop, Yarn, Flume, others
About the presenters

Jonathan Seidman

▪ Software Engineer at Cloudera
▪ Previously Technical Lead on the big data team at Orbitz
▪ Co-founder of the Chicago Hadoop User Group and Chicago Big Data
Enterprise software then…
Enterprise software now…
Well, actually…
Technology Selection
Before we get started…

- A story of two paths:
  - Vendor
  - Customer
There’s a lot on the line

- Limited amount of time to make decisions
- Time and resource cost money
- Don’t want to paint yourself into a corner
- Everyday something new comes up
- Mistakes can be career/project damaging
And many voices

- There are the high profile companies
- There are the venders
- There are the hype trains
- There are the demands from your business
- There is the desire to build up your resume
Gartner Hype Cycle

- Innovation trigger
- Peak of inflated expectations
- Trough of disillusionment
- Plateau of productivity
- Slope of enlightenment

#StrataData
Tricks to Evaluate Hype and Trends

Google Trends
Tricks to Evaluate Hype and Trends

Github activity
Tricks to Evaluate Hype and Trends

Jira Activity

Created vs Resolved Issues Report

This chart shows the number of issues created vs the number of issues resolved in the last 30 days.

Pie Chart Report

Data Table

<table>
<thead>
<tr>
<th>Issue</th>
<th>Issues</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unassigned</td>
<td>316</td>
<td>25%</td>
</tr>
<tr>
<td>Wes McKinney</td>
<td>136</td>
<td>10%</td>
</tr>
<tr>
<td>Lisa L. Kim</td>
<td>125</td>
<td>11%</td>
</tr>
<tr>
<td>Ryan Shue</td>
<td>97</td>
<td>6%</td>
</tr>
<tr>
<td>Derek Maple</td>
<td>81</td>
<td>5%</td>
</tr>
<tr>
<td>Alex Lorenzo</td>
<td>50</td>
<td>2%</td>
</tr>
<tr>
<td>Julien La Cern</td>
<td>27</td>
<td>2%</td>
</tr>
</tbody>
</table>
Tricks to Evaluate Hype and Trends

Meetup.com
Build or buy? It depends

- Internal culture
- Technical religions
- Skill sets
- Opportunity for upside
  - Incremental
  - Revolutionary
  - Control
- Effectiveness of a vendor
- Service Reliability Engineering: Costs and considerations
Consider buy in

- Fear
- Internal buy in
- Forward movement is better than no movement
- Allow for fail fast and restarting
- Build for the Future not a Destination
How to stay focused

▪ Where is the Business Value
▪ All things are possible but not all things are easy
▪ Stay dumb
▪ You must hate the technology before you select it
▪ Treat a vender as a vender not a friend
▪ There must be a desire
Reducing risk

- Interface design
- Fail fast
- Cloud based provisioning
- Containers
- Identify weak points with a passion
Building Successful Teams

Lessons Learned Building Big Data Teams
Build well rounded teams

- Sysadmins
- Developers
- Analysts
- Data Scientists

Other roles:

- Data Protection Officer
- Product Managers
- Network/Systems Engineers
How to find people?
Start with people you already have, but make sure you invest in training…

- Linux, network, DBAs → sysadmins
- Developers → developers
  - Easy if you’re at a company like Orbitz, otherwise maybe not so much.
- Analysts → analysts
- It’s not an easy path though.
  - Set goals instead of micro-managing development.
  - Be prepared to iterate, don’t be afraid to fail.
Also don’t forget other teams

Communication is key

DBAs

Other Project Teams
Hi Jonathan,

If you are interested in the following Hadoop Data Scientist/Administrator position in downtown Chicago, please email your resume and salary requirements to me.

- Develop and extend in-house data toolkits based in Python and Java.
- Consult and educate internal users on Hadoop technologies.
- Improve the performance of financial analytics platforms built around the Hadoop ecosystem.

**Qualifications:**

- 3+ years of experience working with Hadoop 2 (YARN), cluster management experience preferable
- 3+ year of experience with Hadoop SQL interfaces including Hive and Impala
- 2+ years of experience developing solutions using Spark
- Strong systems background, preferably including Linux administration
- Unix scripting experience (bash, tcsh, zsh, python, etc)
- Experience with DevOps tools such as SALT and Puppet as part of a CI/CD development and deployment process.
- Demonstrated ability to troubleshoot and conduct root-cause analysis
Data Scientists

Admins
Think beyond just skills

- Also look for complementary personalities.
- And avoid toxic personalities.
  - But what if they’re really talented?
  - See above.
Customer Engagement

- Your teams should work closely with your customers, whether they’re external or internal.
Thank you!

Ted Malaska | @ted_malaska
Jonathan Seidman | @jseidman