1. **Identify the Use Case** (~1.5 Months)

   - Tackle low hanging fruit first, for a quick win to show fast ROI
   - Choose the project that will provide the biggest potential value
   - Look for the biggest pain in your company/department

2. **Identify the Team** (~2 Months)

<table>
<thead>
<tr>
<th>TEAM</th>
<th>NAME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision Maker</td>
<td>Executive Sponsor / Budget Owner</td>
</tr>
<tr>
<td>Line of Business</td>
<td>Project Manager</td>
</tr>
<tr>
<td></td>
<td>Data Owner</td>
</tr>
<tr>
<td></td>
<td>Analyst</td>
</tr>
<tr>
<td></td>
<td>Subject Matter Expert</td>
</tr>
<tr>
<td>Tech Team</td>
<td>Data Center / Hardware Lead</td>
</tr>
<tr>
<td></td>
<td>Network Administrator</td>
</tr>
<tr>
<td></td>
<td>Hadoop Administrator</td>
</tr>
<tr>
<td></td>
<td>Security Team</td>
</tr>
<tr>
<td></td>
<td>Operations/Monitoring Team</td>
</tr>
<tr>
<td></td>
<td>Software Developers (only if building)</td>
</tr>
<tr>
<td></td>
<td>QA Engineers (only if building)</td>
</tr>
</tbody>
</table>

3. **Identify the Data Sources You Want to Use** (2 weeks)

   - Make sure you have a big data problem. If you have any two of the three variety, velocity or volume of data questions, you have a big data problem.

   **QUESTIONS TO ASK YOURSELF:**

   - What data do I have available?
   - What's missing that I’d like to use?
   - What data elements can I use to link this data with existing data?
   - Do I need to start collecting / storing the data?
   - Who owns the data?
   - Can I have access to it?
   - Are there any security concerns? *e.g. Personal Identifiable Information?*
   - Do security / regulations permit putting this data into the public cloud?
   - How can I get to the data? Push? Pull? Direct access?
   - Load schedule?
   - How much data is it per time unit?
   - How much data do I plan to use each month?
   - For how long?
4. Size the Hardware (1 Month+)

TIPS
- Start with a small cluster. You can always buy more later.
- Consider using a cloud provider like Amazon/Rackspace for fast provisioning
- Time to insight matters more than time per query!
- Hadoop does 3x data replication, no RAID needed.
- Hadoop stores intermediate results - leave twice as much storage headroom as you want to analyze

QUESTIONS TO ASK YOURSELF:
- Do I have an I/O-heavy or CPU-heavy workload?
- Do I really need to evaluate the Hadoop Distribution vendors?
- What are my hardware and software requirements? (see separate Technical Requirements handout)

5. Decide to Build or Buy Your Analytics Solution

BUILD: (~12 Months +)
- Are you ready to build it yourself?
- Do you have a team?
- Does the team have the skill set to build the desired solution?
- Does your organization have the time to hire the right people?
- How much $ and time will it take to train everyone?
- Can you afford to wait 1 year before you get insights?
- Do you have the money for engineering?
- Have you considered ongoing Software Maintenance costs?

BUY: (~3 Months+)
- Have you defined your decision criteria? (See “Succeeding With Big Data Analytics” handout)
- Have you qualified the solution solves your objectives and point points?
- Have you validated your choice? Consider reference calls over POC’s.
- Have you made reference calls?
- Do you have the right people in place to deploy and use this solution?
- Will you need to train people?

6. Determine the Budget

QUESTIONS TO ASK YOURSELF:
- What’s my total cost of ownership? (Contact local rep to discuss)
- Have I considered and budgeted for:
  - People, Consulting
  - Software Acquisition, Support
  - Hardware (Server, network)
  - Logistics (Hardware / people)
  - Operations/Data Center cost (power, cooling etc)
7. Sell Your Project Internally (~2 weeks)
- Identify the sponsor
- Build a business case
- Define your metrics for success
- Evangelize big data in business terms
- Identify and communicate the risk of not doing big data
- Show that the competition is already doing it
- Start small, show the results fast
- Communicate risks clearly
- Demonstrate the use case through pilots
- Get the right people in the room for the pilot
- Involve the consumers of data, not just producers, early

8. Put Your Plan In Action (~1 Month)
- Know the procurement process and major stakeholders
- Buy hardware
- Buy software

9. Deploy
- Deploy hardware
- Install software
- Deploy network
- Integrate monitoring
- Integrate security
- Integrate data

10. Aggregate Your Data
- Identify data source types (connections)
- Are there any firewalls? (verify access to external systems)
- Decide if it’s push vs pull
- Consider performance
- Consider resource utilization and scheduling
- Security
- Data masking / Anonymization
- External Automation
- Data retention
- Data management strategy

11. Train, Train, Train
- Set up workshops instead of lectures
- Educate and empower users on their own data and systems
- Encourage users to think outside the box
- Enable users to access external data sources
- Inspire them to be creative with data discovery

12. Create Your Hypotheses
- TIPS
  - Your Subject Matter Expert knows their data best - NOT your IT team
  - Don’t exclude hypotheses because the data isn’t currently isn’t available
  - Combine internal and external Data sources (e.g. Geo Location + IP, census data + Zip Code)

13. Finding the Insights
- TIPS
  - Give your team self-service tools to enable the subject matter experts
  - Iterate quickly with small samples of the data - then run it against all the data

14. Visualize the Results
- TIPS
  - Use different visualizations to allow for different perspectives
  - Collaborate visually, discuss findings with Subject Matter Experts

QUESTIONS TO ASK YOURSELF:
- What’s the simplest way to get started?
- Will filtering help increase the signal strength?
- Will joining other data increase the signal strength?
Scoping Out Your Big Data Analytics Project

15. Sharpen Insights
- Segment your data and find the outliers first
- Try to identify the reason for the outlier by first segmenting data on available dimensions like geographical area, customer age etc.
- When dimensions aren’t available, enrich data with public data sets like Geo to IP
- Consider there are time and behavior events that depend on each other

16. Interpret
- Identify the root cause of the insights and tie it back to reality
- Visualize and communicate results (infographics)
- Speak in terms of business value (infographics)
- Validate your insights with the business team

17. Put it in Production
- Convert your discovery into scheduled reports and run frequently
- Measure your improvements
- Continuously monitor results
- Share and collaborate

18. Measure ROI
- Present short term ROI to business
- Measure ROI over long term and regularly

19. Disaster Recovery
- How mission critical is it?
- Is backup and recovery sufficient?
- Is cold failover sufficient?
- Do I really need hot failover?

TIPS
- • Measure ROI from the beginning
- • Present short term ROI to business
- • Measure ROI over long term and regularly

20. Iterate
- Iterate on existing use cases and find other insights that can improve the business
- Identify more use cases, never boil the ocean