Executive Summary

As big data technologies move into the mainstream, more and more enterprises are starting to implement their own big data projects.

But before you can have a successful big data project, you have to have a successful big data Proof of Concept. So we’re compiling all the best practice, lessons learned, and advice we’ve gathered over the years to create a practical guide to running a successful POC.

With one important feature: the key to POC success is proving value in a short amount of time. Our workbook aims to help you complete your POC in just six weeks. It’s an ambitious but feasible timeframe. And it’s crucial if you’re trying to prove your big data project can cut it in a real-world enterprise.

As we develop the workbook, we’d like to invite you to review the outline below and tell us if we’re addressing the issues that are important to you. We want the workbook to be absolutely useful for you.

Click the link below to leave your feedback and to pre-register for the workbook, which will be published early 2016.

Leave Feedback
Introduction: The Importance of Starting Small

At this point there’s little doubt that big data represents real and tangible value for enterprises. But the journey from idea to operation is a tricky one. A phased approach to big data isn’t just about making projects more manageable. It’s about ensuring they can adequately prove the value of big data to the enterprise.

And without a well-planned proof of concept to get the ball rolling, it’s far too easy for enterprises to miss out on big data’s true potential.

We’ve written this workbook to help you run a POC that:
1. Is small enough that it can be completed in six weeks.
2. Is important enough that it can prove the value of big data to the rest of the organization.
3. Is intelligently designed so you can grow the project incrementally.

Part I. General Lessons

1. Why POCs Fail

Far too often, POCs run over budget, behind schedule or fail to live up to considerable expectations. So before we dive in to what makes a POC succeed, let’s look at some of the common reasons for failure.

- **A lack of alignment with executives:** Either by going beyond the scope or by promising too much, POCs fail when the executives backing them aren’t involved and their expectations aren’t managed.

- **Lack of planning and design:** You don’t want to spend three of your six weeks just reconfiguring your Hadoop cluster and architecting an enterprise-wide Hadoop-based data lake.

- **Scope creep:** As excitement around the project grows, it’s common for people to add requirements and alter priorities.

- **Ignoring data management:** It’s easy to forget the basic principles of data management and overestimate the tools in the Hadoop ecosystem.

2. Big Data Management Critical Success Factors

Big data management has a crucial role to play in the success of big data initiatives, even at an early stage.

- **Abstraction matters:** The key to scale is leveraging all available hardware platforms and production artifacts (e.g. existing data pipelines). So when it comes to data management, it’s important for the logic and metadata to be abstracted away from the execution platform. This will help preserve all the work you put into a project as systems and technologies change.

- **Leverage your existing talent:** Rather than trying to invest in outside talent for every new skill, it helps to leverage your existing data management expertise to start sooner, keep costs down, and best practices consistent.

- **Leverage automation to operationalize:** Data management in a POC is hard. It may appear you can finish a POC with a small team of experts writing a lot of hand coding. But it really only proves you can get Hadoop to work in a POC environment. It is no guarantee that what you
built will continue to work in a production data center, comply with industry regulations, scale to production and future workloads, or be maintainable as technologies change and new data is onboarded. The key to a successful POC is appreciating what comes next. At this stage you need automation to scale your project. So make the future scale of your initiative part of your planning and design decisions now.

- **Metadata matters:** Repositories of metadata, rules and logic make it a whole lot easier to use a rationalized set of integration patterns and transformations in a repeatable way.
- **Data quality matters:** Data needs to be fit-for-purpose. The only thing worse than no data is bad data.
- **Start thinking about security now:** Security at scale is a whole other ball game when it comes to big data projects. You don’t want your Hadoop cluster to be a free-for-all at the point of operationalization that poses a security risk.

Part II. Picking the First Use Case

1. Limiting the Scope of Your POC

The real challenge around picking the first use case lies in finding something that’s important enough that the business will value it, and small enough that you can solve the problem in six weeks with a low starting cost.

So once you’ve identified a serious business challenge, your next step should be to minimize the scope of the project. The smaller the better. You can limit your scope along these dimensions:

- **The business unit:** In a six week POC, it’s smarter to serve one business unit than it is to try and serve the whole enterprise. Additionally, the smaller the team of business stakeholders the easier it’ll be to meet and manage expectations.

- **The executive:** Executive support is crucial. But trying to cater to the needs of three executives takes much-needed focus and resources away from your six-week timeframe.

- **The team – form, storm, and norm:** Identify champions in business and IT to forge collaborative partnerships of mixed personas. Include only the most essential team members such as subject-matter experts, data scientists, architects, data engineers, data stewards, and DevOps.

- **The data sources:** Aim for as few data sources as possible. Go back and forth between what needs to be achieved and what’s possible in six weeks till you have to manage only a few sources. You can always add new ones once you’ve proved the value.

- **The data:** Focus only on the dimensions and measures that actually matter to your specific POC. If web analytics data from Adobe Analytics (previously SiteCatalyst) has 400 fields but you only need 10 of them, you can rationalize your efforts by focusing on those 10 only.

- **Cluster size:** Typically, your starting cluster should only be as big as 5-10 nodes. However, you should work with your Hadoop vendor to appropriately size your cluster for the POC. Hadoop’s already proven its linear scalability so focus on solving core challenges related to your business use-case for your six-week time frame.
2. Sample Use Cases
The following use cases are from POCs that successfully proved value and limited their scope in a way that ensured they could be added to once the POC was complete.

i. Business Use Cases
- Using general ledger data for real-time fraud processing
- Analyzing website visitor behavior with clickstream data
- Processing a subset of customer data for a customer 360 initiative

ii. Technical Use Cases
In some cases, IT teams have the resources and authority to invest in technical use cases for big data. This may be to identify technical challenges and possibilities around building an enterprise data lake or analytics cloud, for instance. Some of the most common ones are:
- Building a staging environment to offload data storage and ETL workloads from a data warehouse
- Extending your data warehouse with a Hadoop-based ODS (operational data store)
- Building a centralized data lake that stores all enterprise data required for big data analytic projects

Part III. Proving Value
1. Defining the Right Metrics
For your POC to succeed in six weeks, you need to translate expectations of success into clear metrics for success.

Based on your first use case, list out your metrics below:

Quantitative measures:
- E.g. Marketing campaign uplift
- E.g. Impact on sales pipeline
- E.g. Increase lead conversion rates
- …
- …

Qualitative measures:
- E.g. Agility
- E.g. Transparency
- E.g. Productivity and efficiency
- …
- …
1. The Three Layers of a Big Data POC
Big data projects rely on a foundation of three crucial layers, even though only two of the layers get most of the attention.

Visualization and analysis
This includes self-service analytics and visualization tools like Tableau and Qlik.

Big data management
This includes the technology needed to integrate, govern, and secure big data such as pre-built connectors and transformations, data quality, data lineage, and data masking.

Storage persistence layer
This includes persistence technologies and distributed frameworks like Hadoop.

Should Big Data POCs Be Done On-Premise or in the Cloud?
Evaluating the pros and cons of both approaches and explaining when it makes sense to run a Hadoop cluster on your own hardware.

2. The Three Pillars of Big Data Management
i. Big Data Integration
To deliver high-throughput ingestion and at-scale processing for analysts. It’s crucial to:

- Speed up development and maintenance with pre-build transformations and design pattern templates
- Handle all types of data including those with dynamic schemas
- Optimize performance across platforms
- And make it easier to add new sources

ii. Big Data Quality (and Governance)
To manage data at-scale and ensure different users get the data quality they need. It’s crucial to:

- Detect anomalies and make data quality rules scalable
- Manage technical and operational metadata
- Match and link entities from disparate data sets
- Provide end-to-end lineage (for auditability and compliance)
iii. **Big data security**

Data security is usually a checkbox during POCs. But the second your project succeeds, security becomes a whole other ball game.

Foundational security is necessary to:
- Monitor sensitive data wherever it lives and wherever its used
- Carry out risk assessments and alert stakeholders of potential security risks
- Mask data in different environments to de-identify sensitive data

3. **A Checklist for Big Data POCs**

i. **Visualization and Analytics**

1. List the users (and their job roles) who will be working with this layer
2. If you’re aiming for self-service discovery, which tools will your users need (e.g. Self-service data prep)
3. If you’re aiming for reporting or predictive analytics, which tools will you be deploying?
4. Which format will you be delivering data in? (Review sample data in the required format)
5. …

ii. **Big Data Integration**

1. Will you be ingesting real-time streaming data?
2. Which source systems are you ingesting from?
3. Outline the process for accessing data from those systems (who do you need to speak to? Will it be a flat file dump?)
4. How many integration patterns will you need for your use case? (Try to get this number down to 2-3)
5. How many transformations will your use case call for?
6. What type of files are you ingesting? (Review sample data in the required format)
7. How many dimensions and measures are necessary for your six week timeframe?
8. How much data is needed for your use case?
9. …

iii. **Big Data Quality and Governance**

1. Will your end-users need business-intelligence levels of data quality?
2. Or will they be able to work with only a few modifications to the raw data and then prepare the data themselves?
3. Will you need to record the provenance of the data for either documentation, auditing, or security needs?
4. What system can you use for metadata management?
5. Can you track and record transformations as workflows?
6. Will your first use case require data stewards? If so, list out their names and job titles.
7. …

**Documentation**

Even if your POC is successful, you’ll need to be able to share the details of it to both expand the scope of your project and socialize your new capabilities.

As such, it's essential that you document the following:

- The goals of your POC
- The success of your POC as defined by clear metrics for success
- How data flows between different systems
- Your reference architecture and explanations of design choices
- Development support
About Informatica
Informatica is a leading independent software provider focused on delivering transformative innovation for the future of all things data. Organizations around the world rely on Informatica to realize their information potential and drive top business imperatives. More than 5,800 enterprises depend on Informatica to fully leverage their information assets residing on-premise, in the Cloud and on the internet, including social networks.

iv. Big Data Security
1. Will you need different data masking rules for different environments (e.g. development and production)?
2. Which tools will you use for encryption?
3. Have you set up profiles for Kerberos authentication?
4. Have you set your users up to only access data within your network?
5. Do you need application security policies to work within your POC environment?
6. …

v. Big Data Storage Persistence Layer
1. Will you be using Hadoop for pre-processing before moving data to your data warehouse?
2. Or will you be using Hadoop to both store and analyse large volumes of unstructured data?
3. …

4. Reference Architectures for Big Data Management

Conclusion: Making the Most of Your POC
• Big data is a big opportunity for enterprises.
• Poorly planned POCs aren’t just failures in and of themselves. They’re also a failure to prove the value of big data.
• Use the advice and lessons shared in this workbook to hit the ground running and complete a successful POC in six weeks.
• A successful POC is just the start to a constant, iterative and valuable journey towards innovation and insight.