Executive Summary

Businesses looking to uncover new opportunities for growth and efficiency are quickly realizing two things. First, they need to define a big data strategy that delivers real business impact. Second, after years of hype and uncertainty, there is a clear path to defining such a strategy.

Over the past few years, businesses have been creating big data laboratory and big data factory environments.

But it’s only now, with the maturation of technologies and establishment of best practices, that enterprises can successfully deploy both.

Central to this strategy is a solid foundation of big data management.

We’re compiling actionable best practices, lessons learned, and advice into a practical workbook to help businesses like yours to develop this dual-strategy.

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 as you operationalize big data experiments. We want the workbook to be as useful for you as it is for others.

Please click on the link below to leave your comments and to pre-register for the completed workbook.

We thank you for your comments!
Workbook Outline

Introduction: From Experimentation to Monetization

- Beyond the hype, it’s now clear big data’s an important route to growth and invention for large companies.
- But big data success depends on your ability to get at the data, interrogate and productize data assets.
- That means enterprise architectures must serve two purposes.
- The first is to make data available for analysis in a lab environment.
- The second is to make data production-ready for more specific, strategic projects and products.
- The good news is some early movers have already started building out these dual capabilities.
- But the key is that by relying on common architectural components for both environments, they’ve been able to streamline and consolidate their investments in technology.
- Even better, they’ve been able to move their experiments from lab to factory environments in a repeatable, efficient and intelligent way.
- This workbook aims to share their lessons and insights.

Part I. Riding the Elephant in the Room

1. Confronting Hadoop’s Limitations
   There’s a lot of investment and interest in Hadoop. That’s great. Because Hadoop is a big opportunity to simultaneously reduce data storage & processing costs and allow for scale.

   But it isn’t a silver bullet. In fact, it has some non-trivial limitations:
   - There is a significant shortage of Hadoop (and related big data tech like Pig, Hive, etc.) skills. This makes big data projects risky and expensive.
   - Hadoop ecosystem lacks crucial data management capabilities such as data governance, data quality (DQ), Master Data Management (MDM), metadata management, and data security.
   - Security is still limited as Kerberos encryption and access controls aren’t enough for production-grade environments that need to push and pull data from multiple geographies.
   - The ecosystem is still rapidly evolving and new releases, new technologies, etc. emerge all the time. This isn’t always good news as a direction can quickly turn into a dead end: many who bet on MapReduce are struggling to leverage Spark.
   - At the point of operationalization, staffing, maintaining and managing your environments is a whole other ball game. New realm of costs, and new scale makes early hand-coding expensive and challenging to augment.

2. Why Big Data Management Matters
   Considering the non-trivial limitations of Hadoop, data management has a crucial role to play in the success of big data initiatives.
   - It helps you leverage your existing DM and developer talent in big data environments because everyone can still use the DM tools and interfaces they’re used to. Additionally, self-service data prep tools ensure your analysts and scientists don’t have to learn new skills like Java.
• Data integration, governance and quality, security, and master data are actually even more important when dealing with large volumes of data.
• In a factory environment, encryption and access control isn’t enough security. Depending on requirements, data profiling, data masking and even protection for data-in-motion is necessary.
• Smart metadata management and use of repositories of rules and logic is needed to ensure teams can apply flexible standards regardless of the runtime environment.
• By automating data management you can save time and money at the point of operationalization. Time, because developers don’t have to go through twenty pages of code every time they need to add or change a transformation. Money, because you don’t have to rely on manual labour for constant data management and auditing.

Part II. The Lab, the Factory and the Strategy
(Here, we’ll look at the requirements for a lab environment, a factory environment and the data management requirements to enable both in a repeatable way)

1. Building a Big Data Lab
Innovation starts with experimentation. So before enterprises can really turn big data into operational assets – as data products or even as analytical environments to streamline and inform operations – they need to understand what’s possible and what isn’t.

The good news is that the starting costs of a big data lab are relatively low. And once your analysts can roll their sleeves up and start turning insights into viable actions, a big data lab comfortably pays for itself by fueling increased productivity, new opportunities and smarter products.

A big data lab environment has some specific requirements:

– **People:** The focus is on leveraging the intelligence of analysts and the domain knowledge of data stewards. Central IT needs to ‘get out of the way’.

– **Process:** Here central IT needs to be able to provision ‘good enough’ data for analysts to interrogate and experiment on. That means ensuring mass ingestion of multiple data sources for data discovery and self-service data-prep.

– **Technology:** Analysts rely on search and spreadsheets more than any other tool. So any self-service tools made available to streamline analysis should rely on semantic search, a spreadsheet-based interface equipped with automated profiling and discovery capabilities.

2. Building a Big Data Factory
Operationalizing (and potentially monetizing) data assets and products has a different set of requirements.

– **People:** Here central IT and developers take over. Analysts and stewards are still needed to oversee production and make sure requirements for specific projects are met.

– **Process:** Central IT manage DevOps, production support, and the necessary capital investment to operationalize projects, ensuring data is ingested, cleaned and secured in a reliable way.

– **Technology:** In addition to scalable and flexible data pipelines that turn raw source data into actionable information, security and governance have to be ramped up.
3. Why it Makes Sense to Build a Common Infrastructure
   - Consolidated investment
   - Shared metadata
   - Common set of tools make it easier to add new staff

4. The Three Pillars of Big Data Management
Here's what a common infrastructure needs in order to serve both lab and factory environments in the context of big data.

i. Big Data Integration
   - **Universal connectivity:** High throughput DI for multiple schemas, as well as low-latency for real-time streams
   - **Pre-built tools:** Connectors, transformations and parsers to avoid reinventing the wheel every time analysts need to access new sources
   - **Abstraction:** So that the rules, logic and metadata is abstracted away from the execution platform. The key to flexible and maintainable production deployments is leveraging all available infrastructure and hardware, regardless of runtime environment.
   - **A brokerage model:** Leveraging a hub-based orchestration of data flows to reduce redundant efforts, standardize and govern
   - **Staging:** So that depending on the requirements of different users, only the necessary amount of modifications are made to the data.

ii. Big Data Quality (and Governance)
   - **Automated data quality:** In order to manage the data at scale and maintain consistency, reliability of data being fed to analysts. Give them bad data, they’ll give you bad insight.
   - **Business context:** So that stewards can provision standard business terms and definitions through glossaries, etc.
   - **Easy discovery of exceptions:** Normalcy scorecards and exceptions records management so that stewards don’t have to ‘eyeball’ huge volumes of data across multiple sources.
   - **Data lineage:** To provide both auditability and transparency into and the provenance of data and who’s used it.
   - **Relationship management:** So that the stack can infer and detect relationships between entities and domains at scale.

iii. Big Data Security
   - **Profiling and identification:** a 360 view of sensitive data is crucial for a risk-centric approach that holistically protects your data assets even in the case of a perimeter security breach by de-identifying sensitive data
   - **Risk scorecards and analytics:** To automate the detection of high risk scenarios and exceptions based on modelling and score trends.
   - **Universal protection:** To provide masking, encryption and access control across data types (both live and stored data) and across environments (in production and non-production environments).
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.

– Centralized policy-based security: To deliver necessary security and compliance at scale (e.g. Some privacy laws mandate location and role based data controls)

iv. A Reference Architecture for Big Data Management

Conclusion: Interrogate, Invent, Invest

• Big data experimentation is essential to corporate innovation.

• But a big data lab that doesn’t rapidly implement innovative solutions by streamlining them into a production-grade, factory environment is only half complete.

• By making smart architectural and infrastructural decisions, you can de-risk experimentation and streamline production all at once.

• The key is leveraging the big data management capabilities necessary to maximize investments and interest in new storage technologies.