Datameer®

THE GUIDE TO BIG DATA ANALYTICS

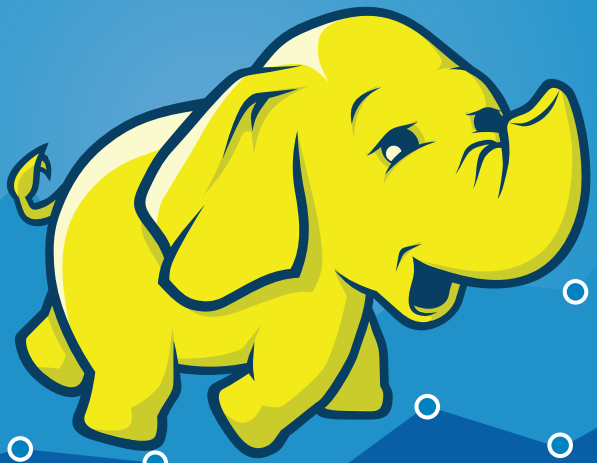


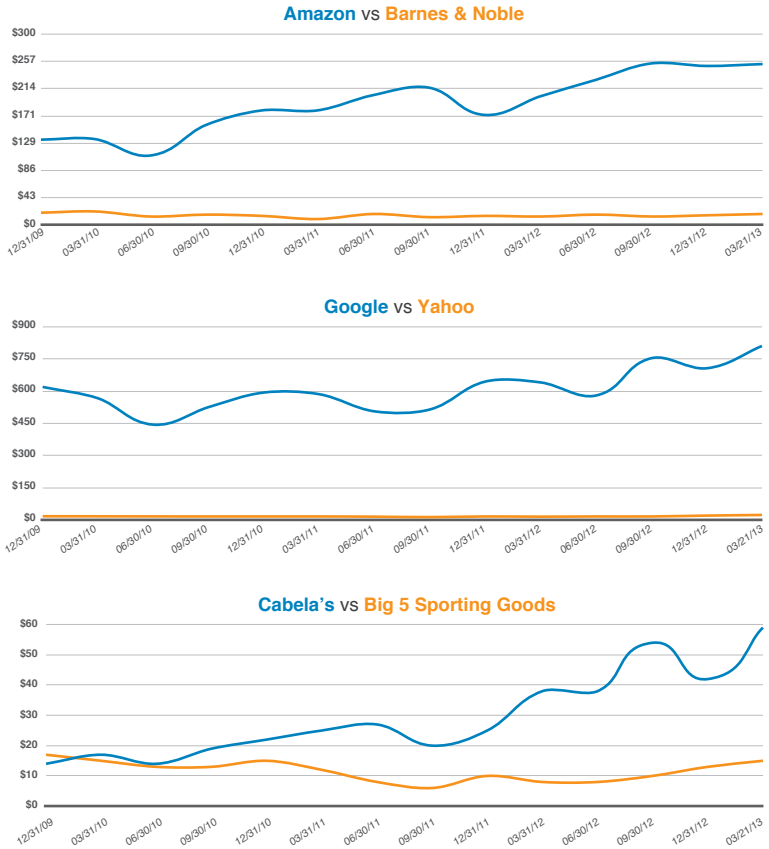
TABLE OF CONTENTS

Why
BIG DATA ANALYTICS?

WHY IS BIG DATA ANALYTICS SO IMPORTANT?

1. Data Drives Performance

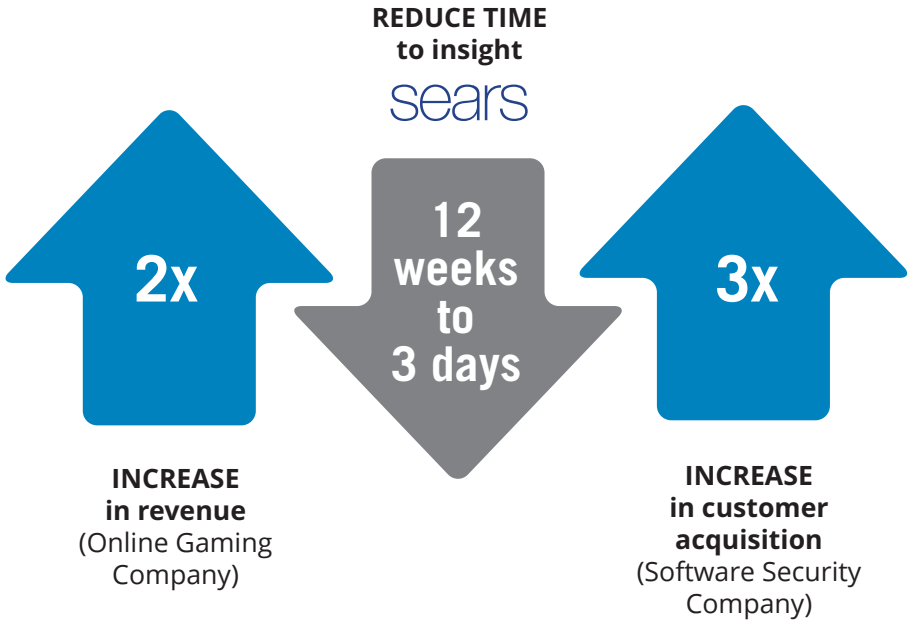
Companies that use data to drive their business (in blue) perform better than companies who do not.



2. Big Data Analytics Drives Results

Companies from all industries use big data analytics to:

- Increase revenue
- Decrease costs
- Increase productivity



** Powered by Datameer*

What is
BIG DATA ANALYTICS?

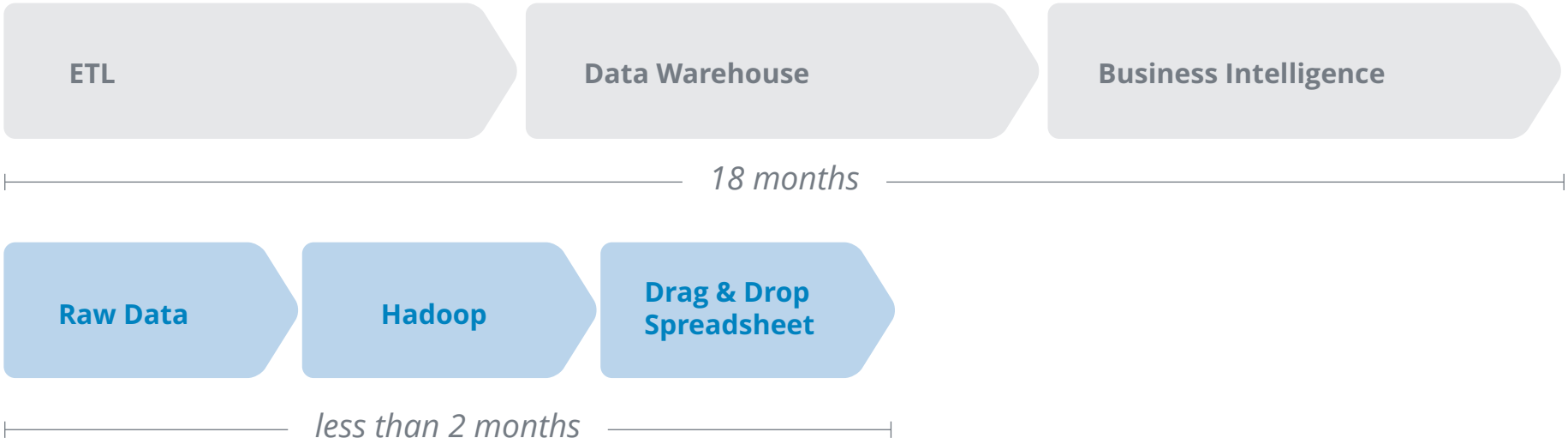
WHAT IS BIG DATA ANALYTICS AND WHAT MAKES IT SO POWERFUL?

The Problem

Before Hadoop, we had limited storage and compute, which led to a long and rigid analytics process (see below). First, IT goes through a lengthy process (often known as ETL) to get every new data source ready to be stored. After getting the data ready, IT puts the data into a database or data warehouse, and into a static data model. The problem with that approach is that IT designs the data model today with the knowledge of yesterday, and you have to hope that it will be good enough for tomorrow. But nobody can predict the perfect schema. Then on top of that you put a business intelligence tool, which because of the static schemas underneath, is optimized to answer KNOWN QUESTIONS.

TDWI says this 3-part process takes 18 months to implement or change. And on average it takes 3 months to integrate a new data source.

The business is telling us they cannot operate at this speed anymore. And there is a better way.



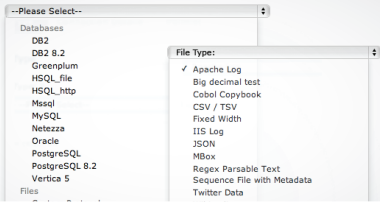
WHAT IS BIG DATA ANALYTICS AND WHAT MAKES IT SO POWERFUL?

The Solution

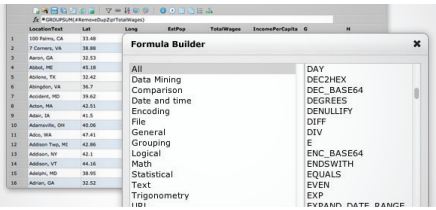
You need a solution with a different approach. First, you need a way to bring in all the different data sources that is much faster than traditional ETL. You need to be able load in raw data. That means loading all metadata in the form it was generated – as a log file, database table, mainframe copybook or a social media stream.

Second, you need a way to discover insights unencumbered by predefined models. You need to be able to iterate with a widely used analysis tool, such as a spreadsheet. Finally, you need a way to visualize those insights and have those visualizations automatically updated. So as your business changes, you can see and get ahead of those changes.

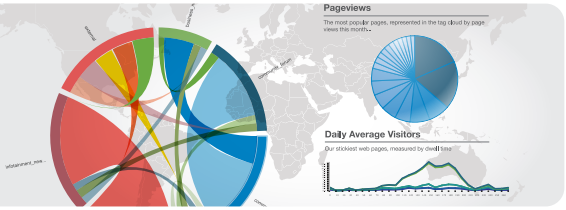
Comprehensive Integration



Schema Free Analytics



Powerful Visualizations



How has

BIG DATA ANALYTICS HELPED?

BIG DATA ANALYTICS SUCCESSES

Use Case	Industry	Benefit
Optimize Funnel Conversion	Software Security	Increased customer conversion by 3x
Behavioral Analytics	Online Gaming	2X Revenue
Customer Segmentation	Financial Services	Decreased customer acquisition costs by 30%
Predictive Support	Enterprise Storage	Decreased customer churn
Market Basket Analysis and Pricing Optimization	Retail	Reduced time to insight from 12 weeks to 3 days
Predict Security Threat	Software Security	Predict security threats within hours
Fraud Detection	Financial Services	Prevented $2B in potential fraud

OPTIMIZE FUNNEL CONVERSION

Identify roadblocks in funnel conversion

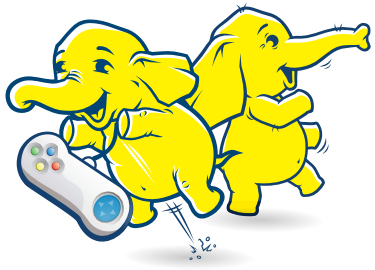
Generating traffic is good. But generating traffic that leads to sales is better. Datameer has helped companies identify which Google AdWords lead To sales. By analyzing Google AdWords, Salesforce, and Marketo data, this company was able to track a lead from AdWord click through to transaction. As a result, this company was able to improve funnel conversion by 3x, leading to $20M in additional revenue.

Customer Conversion Increased by 60%

BEHAVIORAL ANALYTICS

Improve game flow and increase number of paying customers

The game for gaming companies is to increase customer acquisition, retention and monetization. This means getting more users to play, play more often and longer, and pay. First, analysts use Datameer to identify common characteristics of users. As a result, gaming companies can target these users better with the right advertising placement and content. To increase retention, analysts use Datameer to understand what gets a user to play longer. A user who plays longer and interacts with other players makes the overall gaming experience better. To increase monetization, analysts use Datameer to identify the group of users most likely to pay based on common characteristics. As a result of this analysis the company was able to double their revenue to over $100M.



CUSTOMER SEGMENTATION

Target promotions to reduce customer acquisition costs

With mushrooming customer acquisition costs, it has become more important to target promotions effectively. Now, information about a person comes from social media sources in addition to tradition transaction data. To analyze this data, this customer used Datameer to correlate purchase history, profile information and behavior on social media sites. For example, they collected customer profile data. Then they'd correlate this data with transaction history and things the customers "liked" on Facebook. In this way, the company offered a special promotion to a person who liked watching cooking programs and shopped frequently at an organic foods store.

Decrease customer acquisition costs 30%

PREDICTIVE SUPPORT

Identify operational failure and address them before they are reported

A couple of hours of downtime in a store or production environment means lost revenue, sometimes in the millions of dollars. The clues to where downtime may occur are spread across devices around a store or facility including WLAN controllers, mobile devices, routers and firewall devices. For this customer, these devices are used to run operations such as tracking inventory. Each network device generates enormous amounts of machine-generated data. By using Datameer to analyze all network data, this company was able to detect potential network failures faster. As a result, they were able to reduce the number of network failures by 30%.

Reduced network failure by 30%

MARKET BASKET ANALYSIS AND PRICING OPTIMIZATION

Analyze pricing and historical data to price and advertise effectively

In retail, historical inventory, pricing and transaction data are spread across multiple devices and sources. Business users need to pull together this information to understand seasonality of products, come up with competitive pricing, determine which platforms to support so that their online users would have optimal performance, and where to target ads. With Datameer, these business users could do their analysis in 3 days instead of 12 weeks with their traditional tools and heavy IT involvement.

Report time from 12 weeks to 3 days

PREDICT SECURITY THREAT

Identify where security threats may occur

The security landscape is always changing. So changes in behavior can indicate where the next attack may occur. For example, this company used Datameer to follow a virus that started in Russia, moved across Asia, to the US, and forcing Windows upgrades in its path. By seeing where the traffic was generated in particular geographic areas, they could predict where the next security threats would be. This enabled them to proactively go after the security threats and reduce the risk of a data breach, which on average costs organizations $5.5M.

Predict security threats within hours

FRAUD DETECTION

Identify potential fraud

Credit card fraud has changed. Instead of stealing a credit card and using it to buy big ticket items, some credit card thieves have become more sophisticated. For example, they can now making numerous, small transactions that are seemingly benign. But if Joe is making 100 $5 margarita transactions at various locations, something is wrong. By analyzing point of sale, geolocation, authorization, and transaction data with Datameer, this financial customer was able to identify fraud patterns in historical data. This analysis helped the firm identify $2B in fraud. By applying the fraud model to new transactions, the company was able to identify potential fraud and proactively notify customers.

Prevented $2B in fraud

How do I decide

WHETHER TO BUY OR BUILD?

QUESTIONS TO ASK

We have done hundreds of implementation and found that in order to make a decision for build vs. buy, the following questions are critical:

What sort of project is my use case? Data application, discovery or reporting?

Questions to ask when considering the BUILD option:

- [ ] How much time and money will it take to build a team that can build the desired solution?
- [ ] Does your organization have the time to hire the right people?
- [ ] Can you afford to wait 1 year or more before you get insights?
- [ ] Have you considered ongoing software maintenance costs?
- [ ] Are you willing to wait 3 months to add a new data source?
- [ ] Are you willing to manage 3 different teams (for integration, analysis and visualization)?
- [ ] How about all the communication time required to keep in sync regarding changes in the project?

Questions to ask when considering the BUY option:

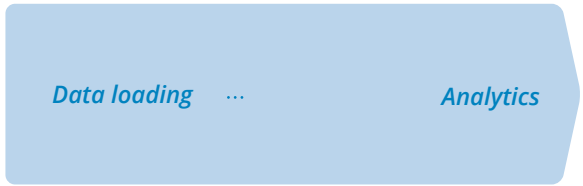
- [ ] Have you defined your decision criteria?
  (See Decision Criteria for selecting a big data analytics solution p.28)
- [ ] Have you confirmed that the solution solves your objectives and proof points?
- [ ] Have you validated your choice? Have you done reference calls?
- [ ] Do you have the right people in place to deploy and use this solution?
- [ ] Will you need to train people?

HOW LONG DOES IT TAKE TO BUILD A BIG DATA ANALYTICS SOLUTION AS COMPARED TO BUYING ONE?

BUILD

14+ months

BUY



3 months

ADVANTAGES OF BUYING VS. BUILDING A BIG DATA ANALYTICS SOLUTION

1 Fewer steps in the process

Instead of hiring developers, ramping them up, defining, developing, testing, deploying and then analyzing, you can go directly to defining and analyzing.

2 Fewer technical requirements

Instead of manually building capabilities for data loading, data parsing, data analytics, visualization, scheduling, dependency management, data synchronization, monitoring API, management UI, and security integration, you can have it in one, purchased platform.

3 Prebuilt integration, analytics and visualization functionality

Seamless Data Integration

- Structured, semi and unstructured
- Pre-built connectors
- Connector plug-in API

Powerful Analytics

- Interactive spreadsheet UI
- Built-in analytic functions
- Macros and function plug-in API

Business Infographics

- Mash up anything, WYSIWYG
- Infographics and dashboards
- Visualization plug-in API

If I Build,

WHAT DO I NEED?

IF I BUILD...

WHAT DO I NEED?

Architecture

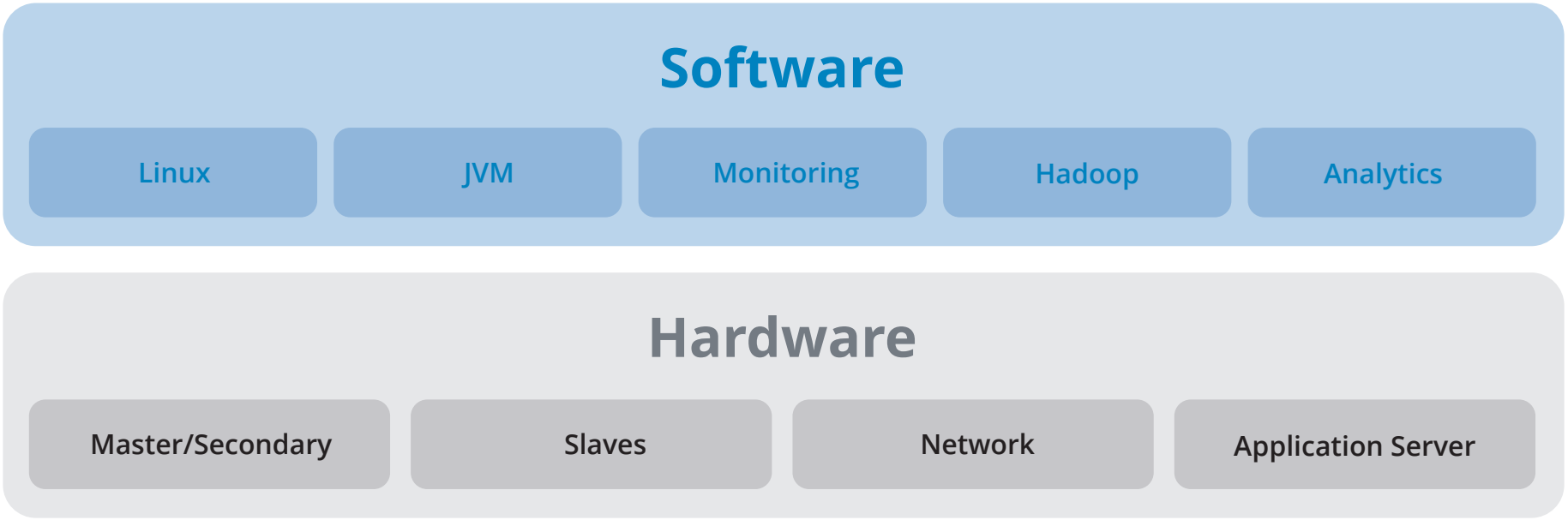
Hardware

Functionality

Process

IF I BUILD...

WHAT ARCHITECTURE DO I NEED?



WHAT FUNCTIONALITY DO I NEED?

In your big data analytics solution, you will need to provide the following:

(see notes section for descriptions)

Data Loading – A software has to be developed to load data from multiple, various data sources. This system needs to deal with the distributed nature of Hadoop on the one side and the non-distributed nature of the data source. The system needs to deal with corrupted records and need to provide monitoring services.

Data Parsing – Most data sources provide data in a certain format that needs to be parsed into the Hadoop system. For example, let's consider parsing a log file into records. Some formats are complicated to parse like JSON where a record can be on many lines of text and not just one line per record.

Data Analytics – In order for data to be properly analyzed, a big data analytics solution needs to support rapid iterations.

Data Visualization – In order for an analyst to see the insights, data needs to be visualized. Integrating visualization is difficult because middleware needs to be built to deliver the data out of Hadoop and into the visualization layer.

Scheduling – All the items discussed above need to be orchestrated and scheduled. Scheduling needs to be easy to configure. In addition, the scheduling needs have monitoring services to notify administrators of jobs that fail.

Dependency Management – There are complex dependencies that must be managed. For example, certain data sets have to be loaded before certain jobs in Hadoop can be run.

Data synchronization – Data often needs to be pushed from Hadoop in to a data store like a database or in-memory system.

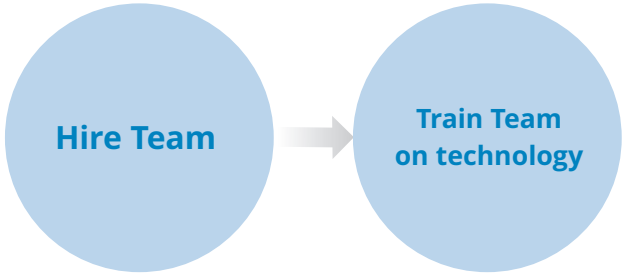
Monitoring API – Every aspect of a big data analytics solution needs to be monitored. Things that need to monitored include who has access to the system, job health, performance, and data throughput.

Management UI – A management user interface is critical for ease of configuration and monitoring.

Security Integration – For security purposes, it is important to be able to integrate with Kerberos and LDAP.

WHAT PROCESS DO I NEED?

In your big data analytics solution, you will need to provide the following:



It is very difficult to hire engineers that are experienced with Hadoop. Most already work for other big companies and/or are very expensive. Hadoop resources can cost $300K/ year or more.

Most often you will find that people have limited Hadoop experience, and you will need to train them on the technology. This is a significant investment in time and money. In addition, practical Hadoop experience only comes from running a Hadoop environment.

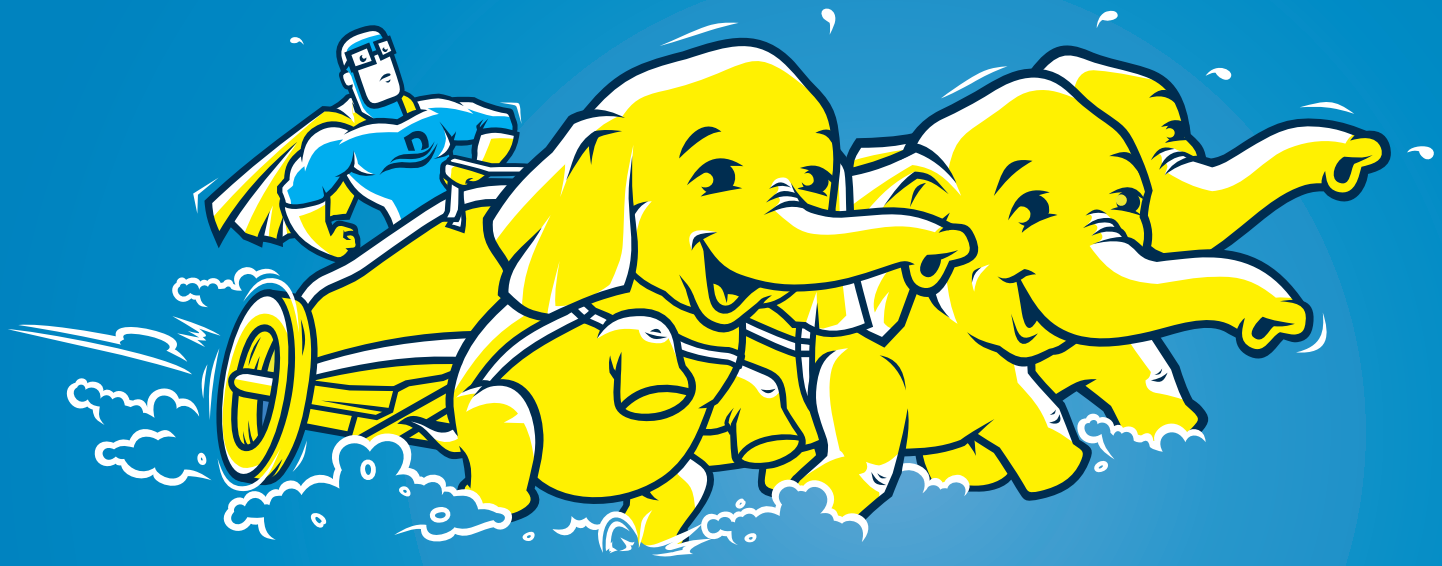
The first step is a design phase and includes requirements gathering (that will change later), technical design, release and resource planning.

In the implementation phase all the “heavy lifting” development work has to be done.

In this phase, as test development is tested, test engineers and a dedicated test environment is required. This often doubles hardware costs.

When building a system frequent deployment into the production system occurs. This means the production system is frequently down or another staging environment is required.

When all design, develop, test and deploy work is done then you can finally start analyzing the data.



How do I Select

THE RIGHT BIG DATA ANALYTICS SOLUTION FOR ME?

DECISION CRITERIA FOR SELECTING A BIG DATA ANALYTICS SOLUTION

After working with hundreds of customers, we've found the following are the top things to look for in a Big Data Analytics Solution.

Ease of use
Does the business need IT to Help?
Can a business analyst use the tool to do analysis?

Data Integration
Does system support native connectors to unstructured and semi-structured data sources (e.g. email, log files, social, SaaS, machine data)? Does it support flexible partitioning of the data so that it is easy to work with large amounts of data? Does the solution support streaming of data so that users will have the most current data? Does the solution provide data quality functions so that the data can be quickly normalized and transformed?

Analytics
Does the solution provide an intuitive environment (e.g. spreadsheet) that business users can quickly use? Does the solution include pre-built analytic functions? Does the system provide a preview to validate analysis and show data lineage for auditing data flows? Do data models have to be defined before insights can be gained? Do analysts need to know what they want to do before they have had a chance to look at the data? Or can analysts look at the data, iterate, make the changes they need and analyze without involving IT?

Visualizations
Does the solution support complete freeform visualization? Or is it just a combination of reports and dashboards?

Integration with existing IT infrastructure
Does the solution support import from and export into other BI systems?

Administration
Does the system support flexible security integration with LDAP and ActiveDirectory?

Extensibility
Does the solution support open API's for custom data connections and custom visualizations?

Architecture
Does the solution run natively on Hadoop? Is a separate cluster required (ideally no separate cluster should be required)? Are there memory constraints, or is the product limited by the availability of the memory within the nodes? The ideal solution should have no memory constraints as that limits the interactivity in analysis and leads to distortion of insights. Does the solution provide a job planner and optimizer to ensure the lowest number of MapReduce jobs is executed?

Vendor requirements
Is the system proven, does it have numerous major releases? How much is it in enterprise use, has it analyze substantial data, (terabytes, petabytes or exabytes)?

Getting Successful

WITH BIG DATA ANALYTICS

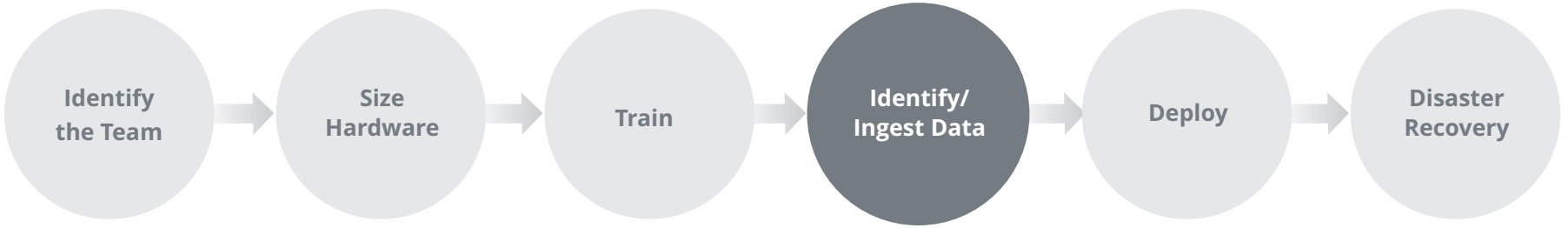
CREATING A SUCCESSFUL PROJECT PLAN

As you go through the steps of creating a successful project plan, IT and the Business must work hand-in-hand.

Here is an overview of the Business and IT best practices. We'll go into the Business best practices in more detail first and then the IT best practices.

Business

IT



BEST PRACTICES
for the Business

IDENTIFY THE USE CASES AND THE TEAM

Identify the Use Cases

Are you trying to:

1. Optimize the funnel?
2. Perform behavioral analytics?
3. Perform customer segmentation?
4. Predict support issues?
5. Analyze credit risk?
6. Perform market basket analysis and optimize pricing?
7. Automate support?
8. Detect and prevent security threats?
9. Detect and prevent fraud?

Identify the Team

Who is the:

1. Budget owner?
2. Project manager?
3. Data owner?
4. Subject matter expert?
5. Business analyst and users?

Data Analyst Best Practices

1. Tackle low hanging fruit first for a quick win to show fast ROI
2. Choose the project that will provide the biggest potential value

BUDGET SO YOU CAN SELL INTERNALLY

To successfully sell internally, answer the following questions:

What's my total cost of ownership?

(Contact marketing@datameer.com for more information on this analysis).

Have I considered and budgeted for:

- People?
- Software Acquisition, Support?
- Hardware (Server, network)?
- Logistics (Hardware / people)?
- Operations/Data Center cost (power, cooling, etc.)?

It is critical to sell your project internally not just at the beginning of the project but continuously.

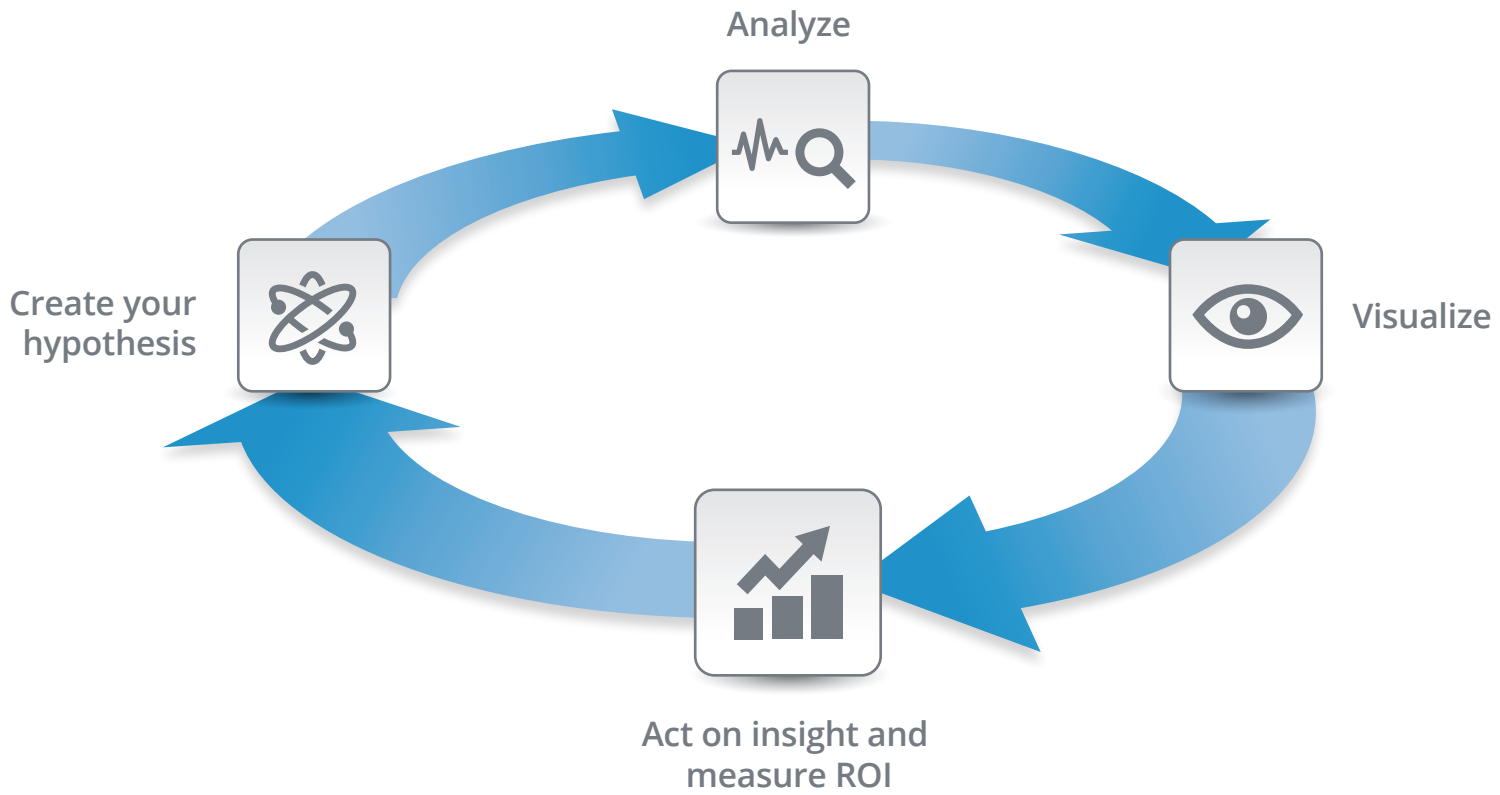
Especially as Big Data Analytics is new and top of mind for many executives and management, selling internally will pay off in the short and long term.

To sell internally, it will be important to:

- [ ] 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
- [ ] Involve the consumers of data, not just producers, early
- [ ]

BIG DATA ANALYTICS IS AN ITERATIVE PROCESS

- Big Data Analytics gets better and more useful the more you use it.
  So iterating has been critical to long term success for our customers.
- Start small and build on your success
- Don't boil the ocean



BEST PRACTICES for IT

IDENTIFY THE HARDWARE YOU NEED

Master/Secondary

- Fault Tolerant
- RAID
- Dual Network Power, etc.

Slaves

- 4+ HDD
- 1 CPU core per HDD
- 2GB RAM/core

Network

- 1GB fully unblocked network switches
- 10GB uplink for Top-of-Rack switches
- Dual Network Power, etc.

Application Server

- 2 HDD's
- 16GB RAM

Questions to ask:

Do I have an I/O or CPU intensive workload?

Do I really need to evaluate the Hadoop Distribution vendors?

Have I considered power consumption?

Have I considered cooling needs?

Have I considered the weight of hardware?

Does our data center support it?

Best Practices:

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

IDENTIFY AND AGGREGATE YOUR DATA SOURCES

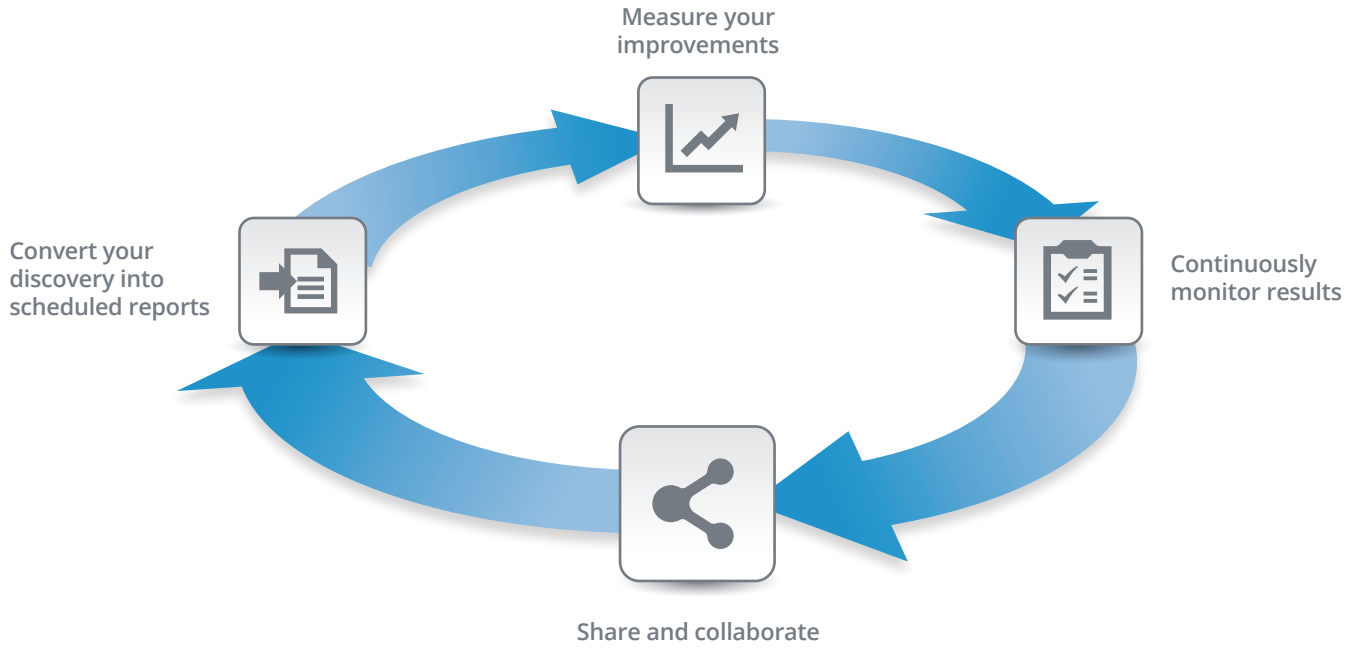
Answering the following will help identify and aggregate the right data sources in particular:

1. What data do I have available?
2. What's missing that I'd like to use?
3. What data elements can I use to link this data with existing data?
4. Will the data be pushed or pulled into Datameer and Hadoop?
5. What is the performance expectation?
6. What type and level of security is needed? Are you looking for data masking or anonymization?
7. What is the data retention policy?
8. What is the data management strategy?

PUT YOUR PLAN IN ACTION & DEPLOY

Now that you have the insights, let's automate the process and make creating these impactful insights repeatable. To do that, the following are important to incorporate into your plan.

1. Convert your discovery into scheduled reports and run as needed
2. Measure your improvements
3. Continuously monitor results
4. Share and collaborate



ENABLEMENT

Enabling your organization, across the different functions and business stakeholders, can be challenging. The following things have helped with the customers we worked with:

1. Set up workshops instead of lectures
2. Educate and empower users on their own data and systems
3. Encourage users to think outside the box
4. Enable users to access external data sources
5. Inspire them to be creative with data discovery

KEY TAKEAWAYS

Land and expand. Start small and go after the low hanging fruit. Then build upon those successes to go broader.

A successful Big Data Analytics solution makes it easier and faster to analyze broader sets of data. You need a solution that:

1 Is faster because it has fewer steps in the process.

Instead of hiring developers, ramping them up, defining, developing, testing, deploying and then analyzing, you can start with defining and analyzing.

2 Has fewer technical requirements

Instead of manually building capabilities for data loading, data parsing, data analytics, data visualization, scheduling, dependency management, data synchronization, monitoring API, management UI, and security integration, you can have it all in one platform. Buying the right solution also means you don't have force fit a new technology into old architecture.

3 Prebuilt integration, analytics, dashboards

Data integration – Quickly ingesting data from many different types of data sources is critical to the success of a big data project. Eliminating the IT intensive ETL and modeling processes is critical to allowing analysts to be able to meaningfully interact with data and to prevent long and expensive waits while data is prepared. To achieve this, a solution should have a large set of native connectors for the integration of all types of disparate data – structured, unstructured and semi-structured (e.g. databases, file systems, social media, mainframe, SaaS applications, XML, JSON).

Analytics – Analysts need a tool to work with the data, turn it into a form that can be usable and then perform the analysis. The means having pre built transformations to clean up the data so that is usable as well as prebuilt analytic functions. These functions should include ones to support unstructured data and sentiment analysis. This tool should also provide a preview to validate analysis and show data lineage for auditing data flows.

Visualization – Beyond static dashboards, analysts need Infographics where data visualizations have no built-in constraints. Users need to be able to drag and top any widget, graphic, text, dashboard or info graphic element as needed or desired.

EXPERTS IN BIG DATA ANALYTICS

You're not alone!

Combining decades of experience in Hadoop, data management and business intelligence, there are people who have been working with customers around the world, across industries such as Communications, Financial Services, Retail, Internet Security, Consumer Packaged Goods, Transportation, Logistics, and deep into use cases such as fraud detection, behavioral analytics, customer segmentation, pricing optimization and credit risk analysis.

We are happy to help, so please reach out to the following experts with any questions or thoughts you might have. Also please send any questions, comments or thoughts on this guide to **khsu@datameer.com**

STEFAN GROSCHUPF
CEO

Stefan Groschupf is the co-founder and CEO of Datameer, the only end-to-end data integration, analysis and visualization platform for big data analytics on Hadoop.

FRANK HENZE
VP, Product Management

Frank Henze is Vice President of Product Management at Datameer with over a decade of experience in building enterprise software systems.

PETER VOSS
CTO

Peter Voss is CTO at Datameer with extensive experience in software engineering and architecture of large-scale data processing.

EDUARDO ROSAS
VP, Services

Eduardo Rosas is VP of Services and heads the implementation of Big Data Analytics projects that triple customer conversion rates, lead to over $20M in additional revenue, and cut IT costs of big data analytics projects in half.

KEN KRUGLER
President, Scale Unlimited

Consulting and training for big data processing and web mining problems, using Hadoop, Cascading, Cassandra and Solr. Apache Software Foundation member, committer for Tika open source content extraction toolkit.

RON BODKIN
CEO, Think Big Analytics

Ron founded Think Big to help companies realize measurable value from Big Data. Previously, Ron was VP Engineering at Quantcast where he led the data science and engineer teams that pioneered the use of Hadoop and NoSQL for batch and real-time decision making.

JOE CASERTA
CEO, Caserta Concepts

Joe Caserta is a veteran solution provider, educator, and President of Caserta Concepts, a specialized data warehousing, business intelligence and big data analytics consulting firm.

PHIL SHELLEY
Former CTO, Sears

Dr. Shelley advises companies on designing, delivering and operating Hadoop-based solutions for Analytics, Mainframe Migration and massive-scale processing. He was CTO at Sears Holdings Corporation (SHC) and lead IT Operations focusing on the modernization of IT across the company.