Taking an Enterprise Wide Approach to Big Data Initiatives

Pete Schrader, *Partner*, PwC
Matt Bonser, *Director*, PwC
Manoj Motiwala, *Manager*, PwC

Professional Techniques – T23

"[Gartner] predicts that **data volume will double over the next two years**" - Gartner

Big Data is everywhere -- every industry publication you read, every conference you go to. But, exactly **what does it mean**, and when you launch your Big Data initiative, **what should you be doing** to enhance your chances of success?

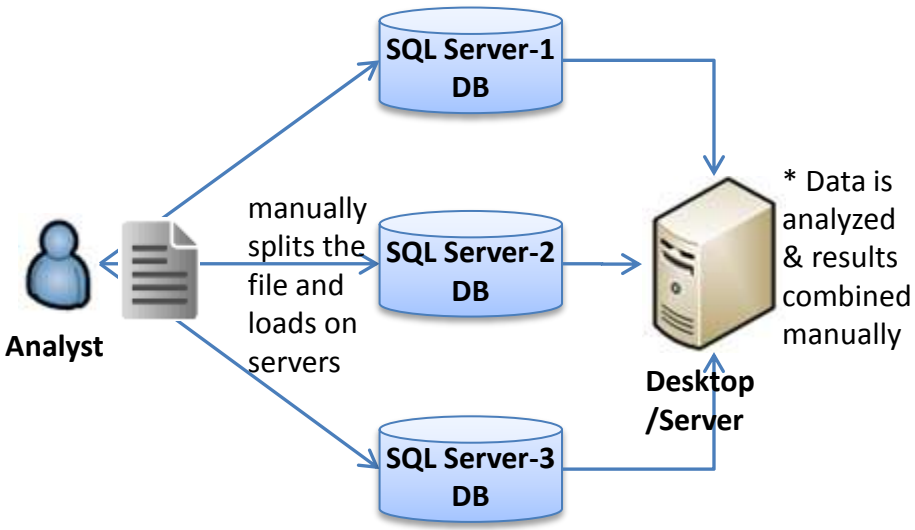
70% of enterprises are either deploying or planning to deploy Big Data solutions within the **next 18 months**.

- IDG Enterprise 2014 Big Data survey

Traditional vs. Big Data Analytics - Example

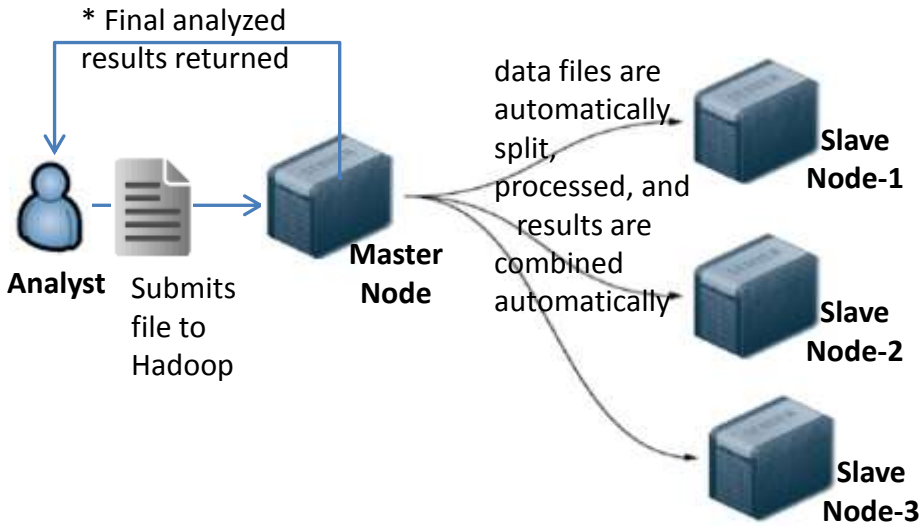
Business Need: Validate aggregated revenue amounts from transaction data
Analytic: Assess completeness, accuracy and recalculate revenue, noting outliers
Data: 1B transactions, 10 TB

Traditional Data Analytics
using Database tools, SAS, SPSS…..



Outcome: 2 weeks to generate results

Big Data Analytics
using Hadoop tools, Hive, R …..



Outcome: 2-3 days to generate results

Agenda

1. What Is Big Data and Why Is It Important?
2. Understanding What You Want From Big Data
3. How You Get to Big Data
4. Risk Considerations
5. Involve the Right People From Across the Business
6. Choose the Right Implementation Approach
7. Enabling the Future With the Choices You Make Now

What Is Big Data and Why Is It Important?

What is Big Data?

"Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization."

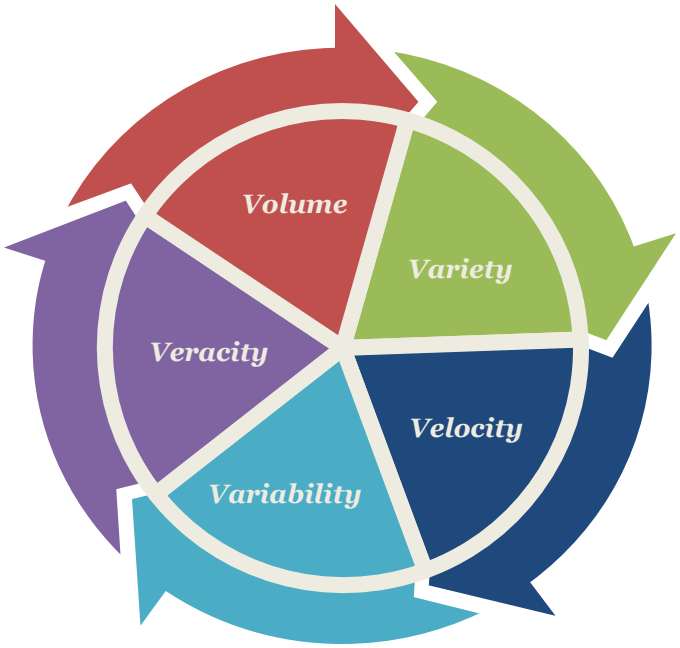
- Gartner, 2012

What is Big Data? - Adding to the Definition

Volume - The volume of available data increases daily as more and more actions are tracked

Velocity/Variability - With volume comes velocity. Data flows can be highly inconsistent with periodic peaks. Data from RFID, Logs, Machine, Social Media etc contribute to this

Variety - Data today comes in all types of formats such as unstructured text documents, email, video, audio, stock ticker data and financial transactions



Veracity- Refers to the biases, noise and abnormality in data. Is the data that is being stored, and mined meaningful to the problem being analyzed

Or, In Real Life Terms…

500 MM	Number of Tweets per day	*http://abcnews.go.com/Business/twitter-ipo-filing-reveals-500-million-tweets-day/story?id=20460493
100 Hours	100 hours of video are uploaded to YouTube every minute	*https://www.youtube.com/yt/press/statistics.html
350 MM	Photos uploaded on Facebook every day	*http://www.businessinsider.com/facebook-350-million-photos-each-day-2013-9
4.5 BN	Likes by Facebook users per day	*https://zephoria.com/social-media/top-15-valuable-facebook-statistics/
183 BN	Emails sent per day	*http://sourcedigit.com/4233-much-email-use-daily-182-9-billion-emails-sentreceived-per-day-worldwide/

Challenges with Traditional Data Analytics

- Inability to efficiently store, process and analyze:
  - Increased data volume from new avenues - campaign analysis, social media, risk/fraud monitoring, devices etc.
  - Real time influx of data - logs, tweets, posts, blogs, machine data
  - Complex semi-structured data, unstructured data and data generated in process
- Expensive implementations; scaling up/down is not a smooth process
  - Adding more processing/servers after a point is not that beneficial
  - Difficult to dis-invest from long term hardware and software costs
- High dependency on network and demands on bandwidth
  - Step 1 - Normalized data is moved to shared file system
  - Step 2 - Data is transported/imported to centralized database or statistical tool
  - Step 3 - Execution of queries requiring data store of output
- Failures during data load are difficult to handle
  - A single failure can disable a process to execute (and other dependent process)
- Slower data processing carried out on single centralized server
  - Inability to automatically “divide and conquer”

Why Big Data Solutions are so Important

- Data Revolution has provided opportunities and challenges
  - Transforming data in to insights requires change
- Ability to store data in all sizes, formats - coming at any frequency
  - Beyond the storage and processing of traditional database systems
- Improved decision making
  - Access to larger sample data or even entire population
  - Models get processed much faster
- Higher Return on Investment (ROI)
  - Clients use Hadoop to store and analyze data for multiple use cases
  - Hadoop is open source and is less expensive then traditional BI solutions
- Validation of data coming from existing and new avenues
  - Is data complete and accurate for the intended use
  - How long is data valid and how long should it be stored
  - Determining authenticity and value associated with data

Understanding What You Want From Big Data

Opportunities with Big Data Analytics

- Inexpensive implementation on commodity hardware
  - Easy to scale up and down without impacting the current processes
- Evolving open source technologies geared towards future demands
  - Process terabytes to petabytes of structured and un-structured data
  - Technologies include: MapReduce, Pig, Hive/Impala/Tez, Talend, Mahout, R, Spark/Drill, Casandra
- Efficiency in storing, processing and analyzing Big Data
  - Massive parallel data storage and management
  - Processing is done where the data is stored
  - No data synchronization is required

Traditional vs. Big Data Analytics

Traditional

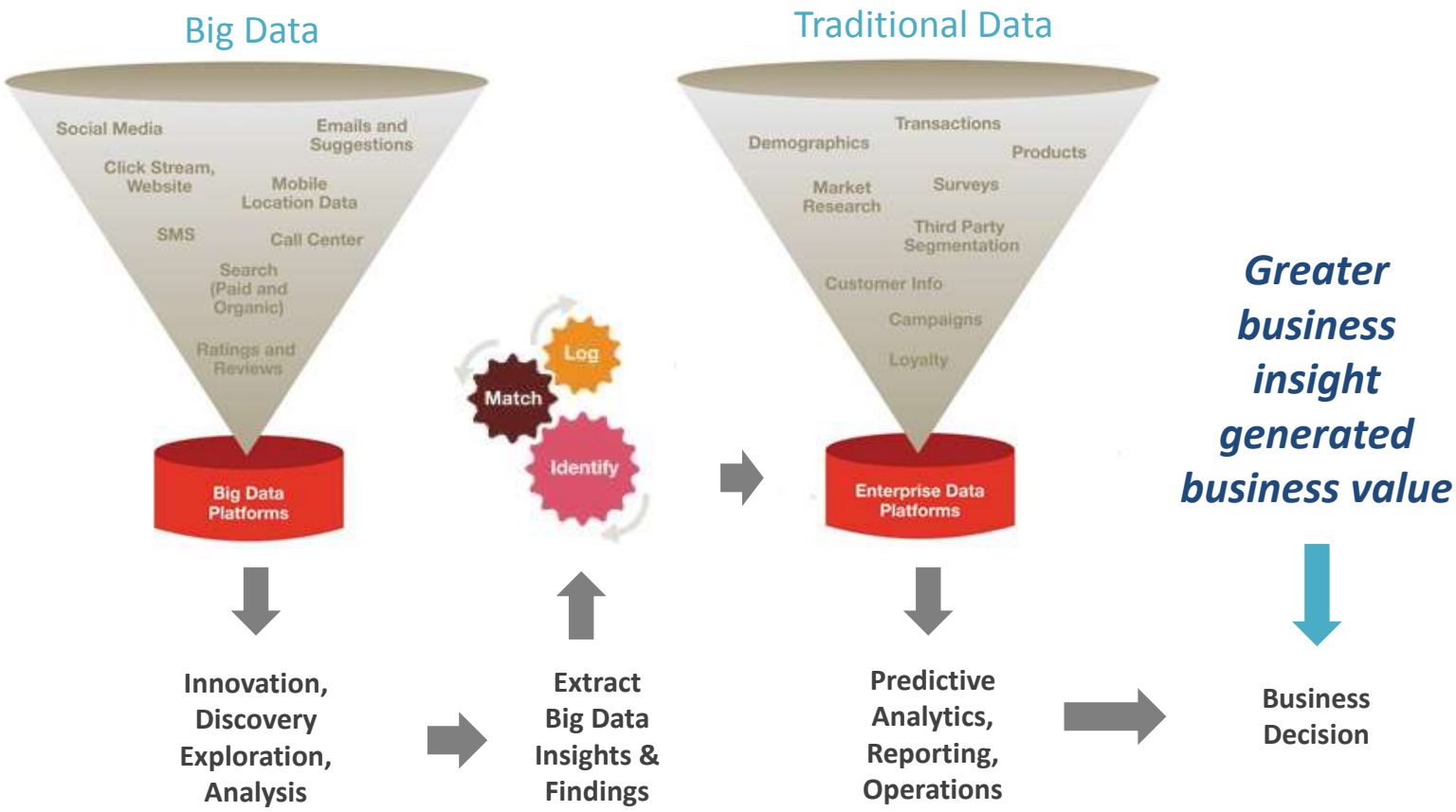
- Built on top of the relational data model
- Data often used is well understood, cleansed, and in line with business metadata
- Traditional analytics is often batch oriented
- Parallelism in a traditional analytics system is achieved through costly hardware like Massively Parallel Processing (MPP)

Big Data

- Big Data consists of structured, semi-structured, and unstructured data
- Unstructured data that is usually stored in columnar databases
- Unstructured data is not well formed or cleansed
- Big Data analytics is aimed at near real time analysis of the data
- While there are appliances in the market for Big Data analytics, it can also be achieved through commodity hardware and new generation of analytical software (e.g., Hadoop)

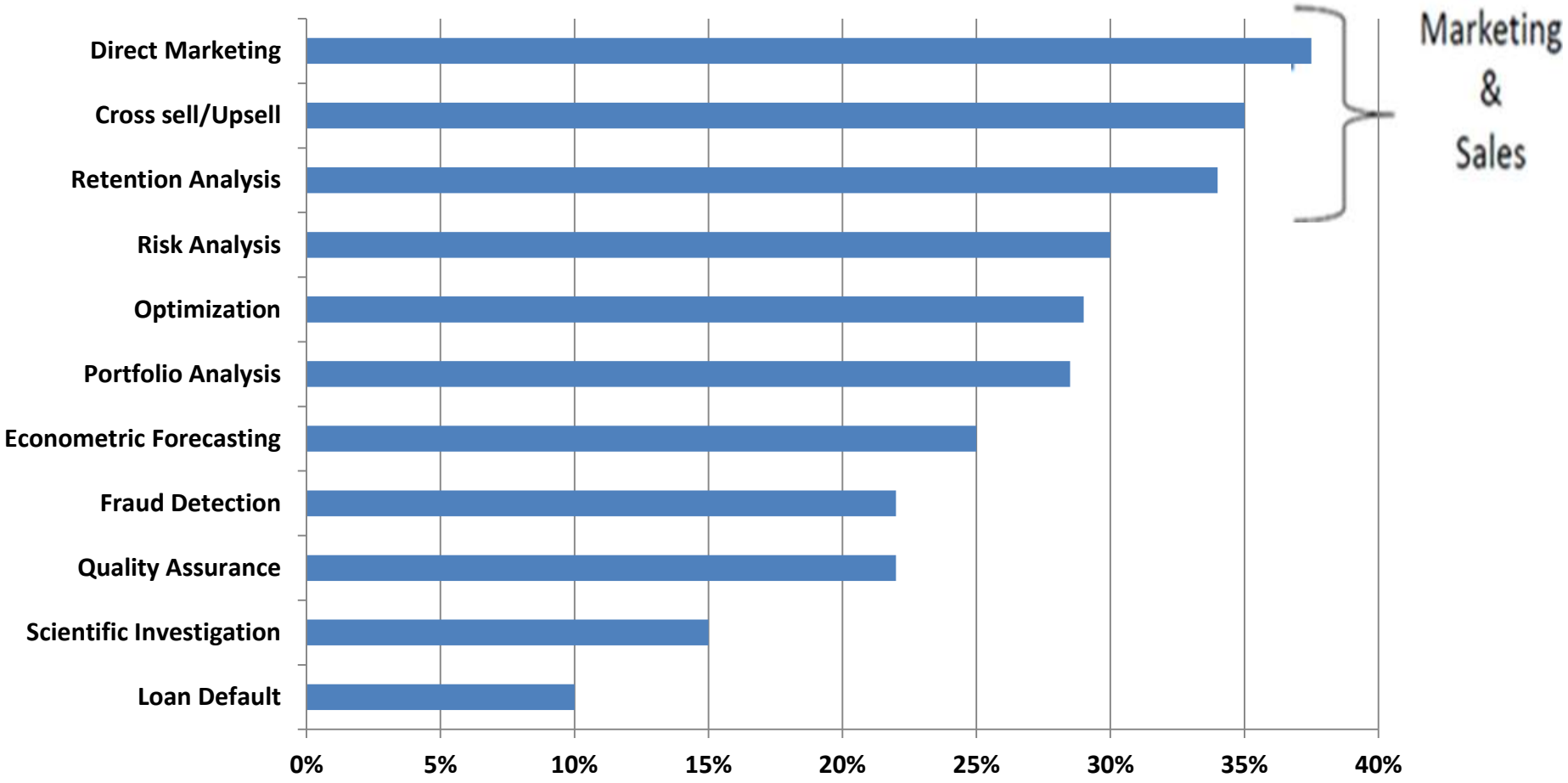
Future of Analytic Data Processing is a Hybrid of Analytical Database & Hadoop

Big Data analysis does not replace other systems. Rather, it supplements other analytic solutions, data warehouses, and database systems essential to financial reporting, sales management, production management, and compliance systems.



What is the Business Problem You are Trying to Solve?

Analytics are being used in companies in following capacities



Source: TDWI, 2013

What is the Business Problem You are Trying to Solve?

Industry use cases for using Big Data

- Improving financial operations
  - Credit/loan risk scoring - decreasing risk of default
  - Fraud and AML* detection - detecting more instances of fraud and AML
  - Fraud discovery - discovering whole new types of frauds
  - Reduce costs - reducing product and operations costs

  * AML = Anti-Money Laundering

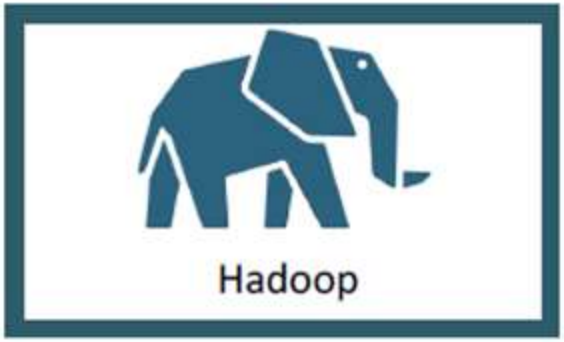
- Improving marketing of interest
  - Customer lead scoring - improve propensity to buy, attain new customers
  - Market segmentation - detect customer types
  - Personalized recommendations - open new cross sell opportunities
  - Churn prevention - identify customers about to churn and how to retain them

- Improving pricing/products of interest
  - Algorithmic pricing - targeted pricing, deciding price points from offer feedback
  - Product design - product targeting, deciding which product features optimize revenue

How You Get To Big Data

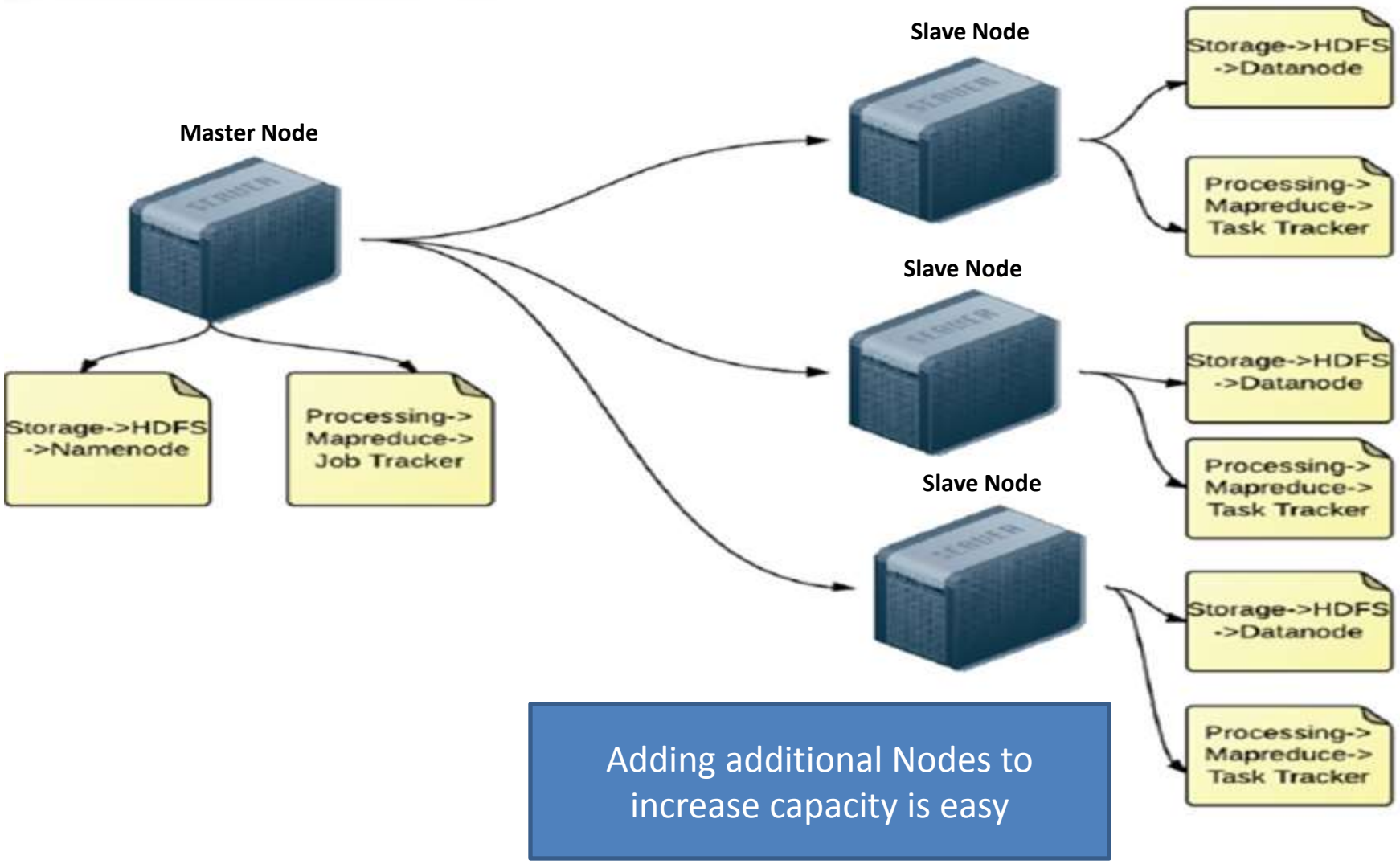
Big Data Landscape

Apache Ambari
http://incubator.apache.org/ambari



Visualizing Hadoop

Master Node automatically distributes data and processing to multiple Slave Nodes



Big Data Tools: Vendor Categories

Knowing the platform requirements drives potential vendor decisions

Hadoop Distribution (Pure Play)

Vendor Products

- Cloudera (CDH)
- Hortonworks
- MapR Technologies
- Apache

Hadoop Analytics

Tools

- MapReduce/Pig/Hive/Talend/Pentaho (data management)
- Impala/Spark/Shark/Tez/Drill (data discovery/analysis)
- HBase/Cassandra/MongoDB (no SQL DB)
- Mahout and 'R' (predictive analytics)
- Tableau/Qlikview/MicroStrategy/Spotfire /Karmasphere (visual analytics)
- IBM BigSheets/Platfora/Datameer/ MS PowerView (spread sheet like analysis)

Integrated Stack

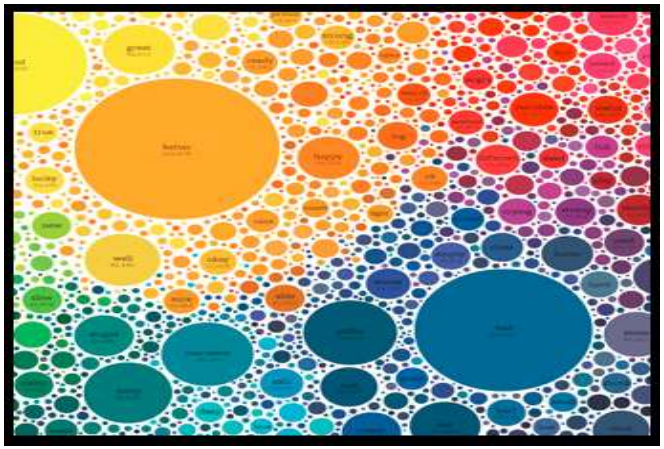
Vendors

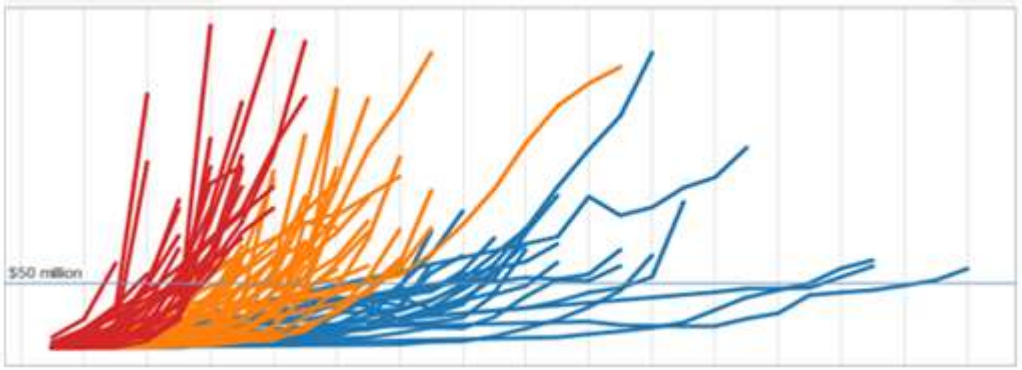
- IBM InfoSphere BigInsights
- SAS
- Oracle Big Data Appliance
- EMC Greenplum HD
- MS SQL Server Stack
- SAP Hana

Big Data Visualization

A picture is worth a thousand words

- Visualization software allows analytical results to be understood more holistically
- Find relevance among the millions of variables, communicate concepts and hypotheses to others, and even predict the future
- Interactive Visualization: Use computers and mobile devices to drill down into charts and graphs for more details





Big Data Visualization

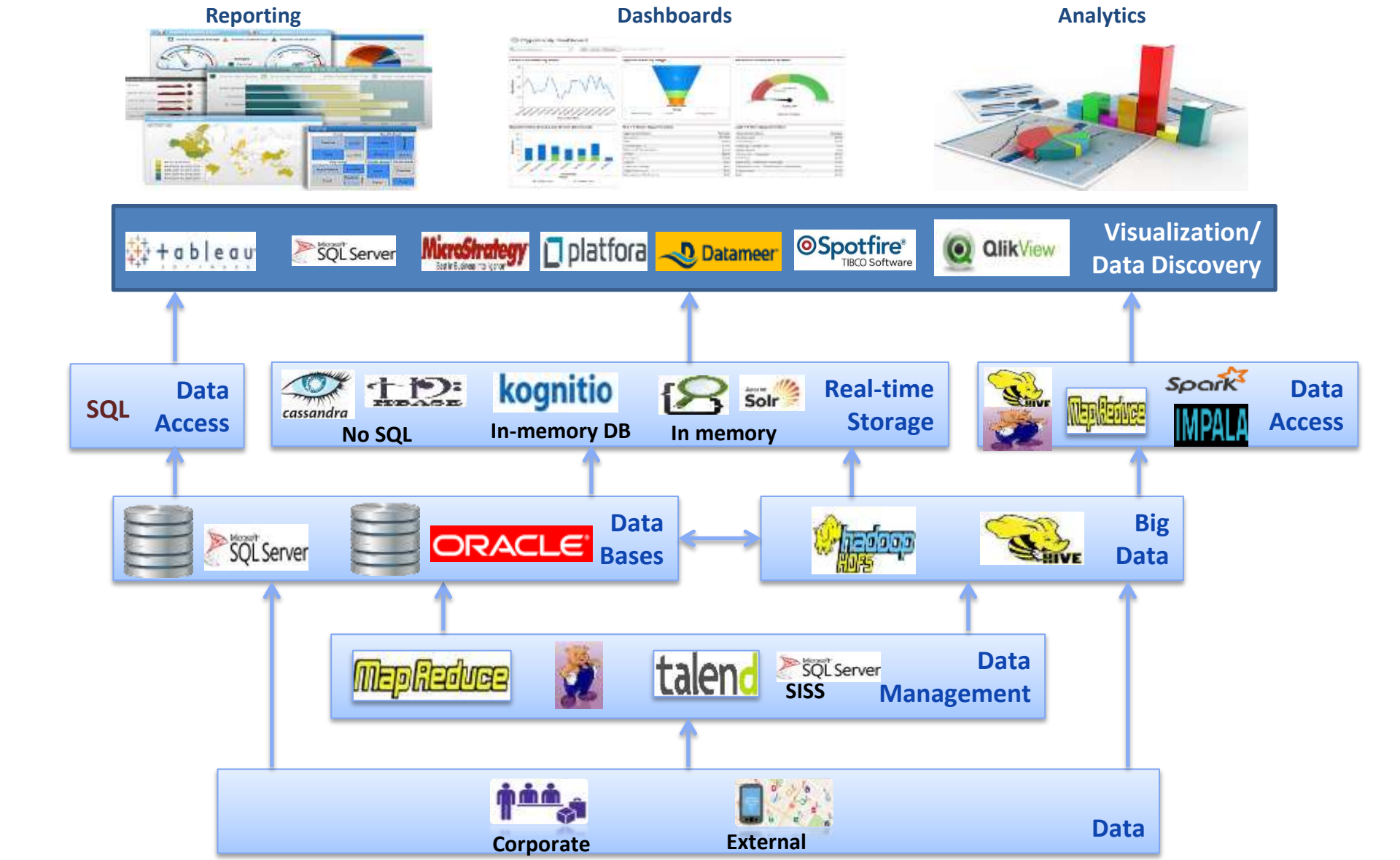
Visual Analytics Tools

- Several companies are marketing tools tailored for Big Data or Hadoop with visual analytics

- Tool features includes:
  - Spreadsheet-like interface with functions
  - Variety of built-in charts, interactive dashboards
  - Connections to data stored in Hadoop that typically require data in Hive/connection driver
  - Interactivity of Hadoop data remains limited by the batch processing nature of Hive

- No one size fits all, be flexible and adaptive, and involve the right people in the decision making process

Big Data - Bringing Technology Together



Risk Considerations

Risk Considerations

- Business Risks: Ensure end goal is defined
  - Avoid lack of alignment from the Business and key stakeholders by involving them throughout the process, including regular communication
  - Avoid a lack of alignment with strategic objectives by establishing a governance process early in the engagement
  - Avoid dissatisfaction from key users by setting expectations for delivery timelines early and communicate any changes in a timely manner
  - Involve the right people in the requirements generation phase to avoid missed items or under/over scoping

Risk Considerations

- Technology Risk: Avoid over/under investing in infrastructure
  - Understand platform requirements to prevent over or under investment in technology risk
  - Develop an understanding of data growth predictions to make sure that the solution will be capable of meeting future needs
  - Involve people from across the IT function to ensure that any technology fits into the overall IT roadmap

Risk Considerations

- Resource Risk: Staffing shortages are a real problem
  - Resources with the right skills are scarce, so start planning early in order to be able to onboard the right people at the right time
  - Existing resources probably won't have the required skill sets so develop a training program to be able to upskill them accordingly to manage both risk to delivery as well as people satisfaction
  - Consider if any level of organizational change management is required to help manage delivery risk

Risk Considerations

- Security & Privacy Risk: Securing all the data is a challenge

  Build up a risk assessment program:
  - Understand what your critical data assets are, and the extent to which they are included in the initiative
  - Consider how data is to be stored/accessed along with any other risks appropriate to your data set
  - Understand your organization's privacy requirements and security framework
  - Determine if critical data assets are being protected in line with requirements and make changes as appropriate

Involve the Right People From Across the Business

Good Governance Structures are Important

- How is the information going to be used, maintained, stored, secured and accessed?
- Big Data solutions are open sourced – they are an evolving solution developed to gain information, not secure the information.
- Architecture
  - Physical location (in house, consultants, cloud)
  - Operating system
  - Tools
- Support maintenance and security
- Sustainability of the effort upfront

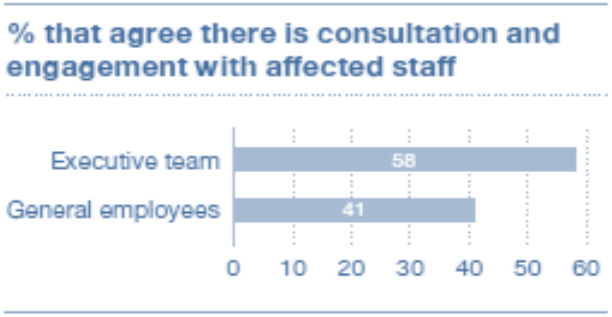
Measure twice, cut once

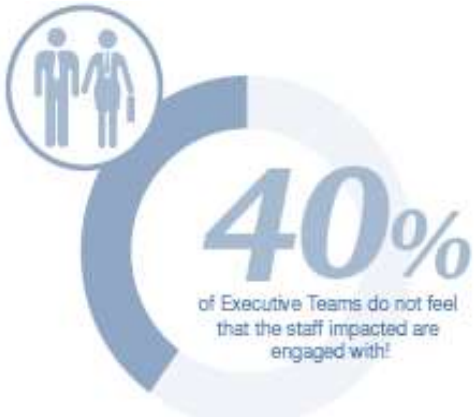
Involve people from across the business, not just technology, as they are ultimately going to be the consumers of the project outputs

Treat this as a whole of business solution

Leadership Buy-In is Important

Misalignment of executive leadership and project teams





PwC, 4th Global Portfolio and Program Management Survey, September 2014

There needs to be a level of understanding that this will be an ever-evolving solution

Choose the Right Implementation Approach

Plan Strategically, Implement Tactically

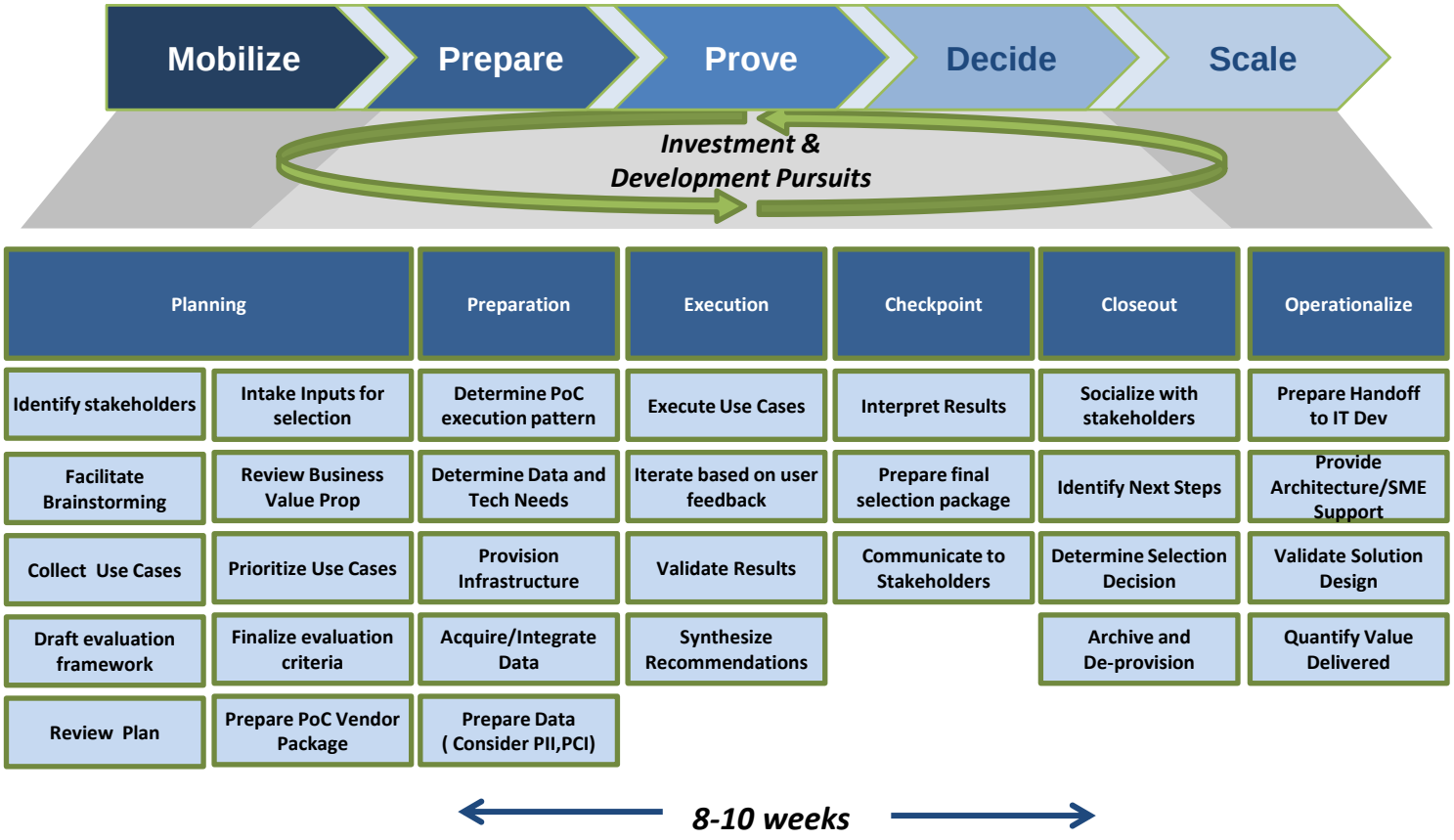
Traditional Project Management and Development Approaches May Not Work

Consider an Agile, or Hybrid Agile Approach

- Importance of iterations and proof of concept
- Co-locate resources
- Regular release of usable code
- Scope management – limited flexibility because complete ecosystem must be built

Vendor/Technology/Application Selection: Activities/Timeline

Big Data solution selection should be driven by a sound information strategy and executed as a collaboration between Business and IT stakeholders over a typical period of 8-10 weeks.



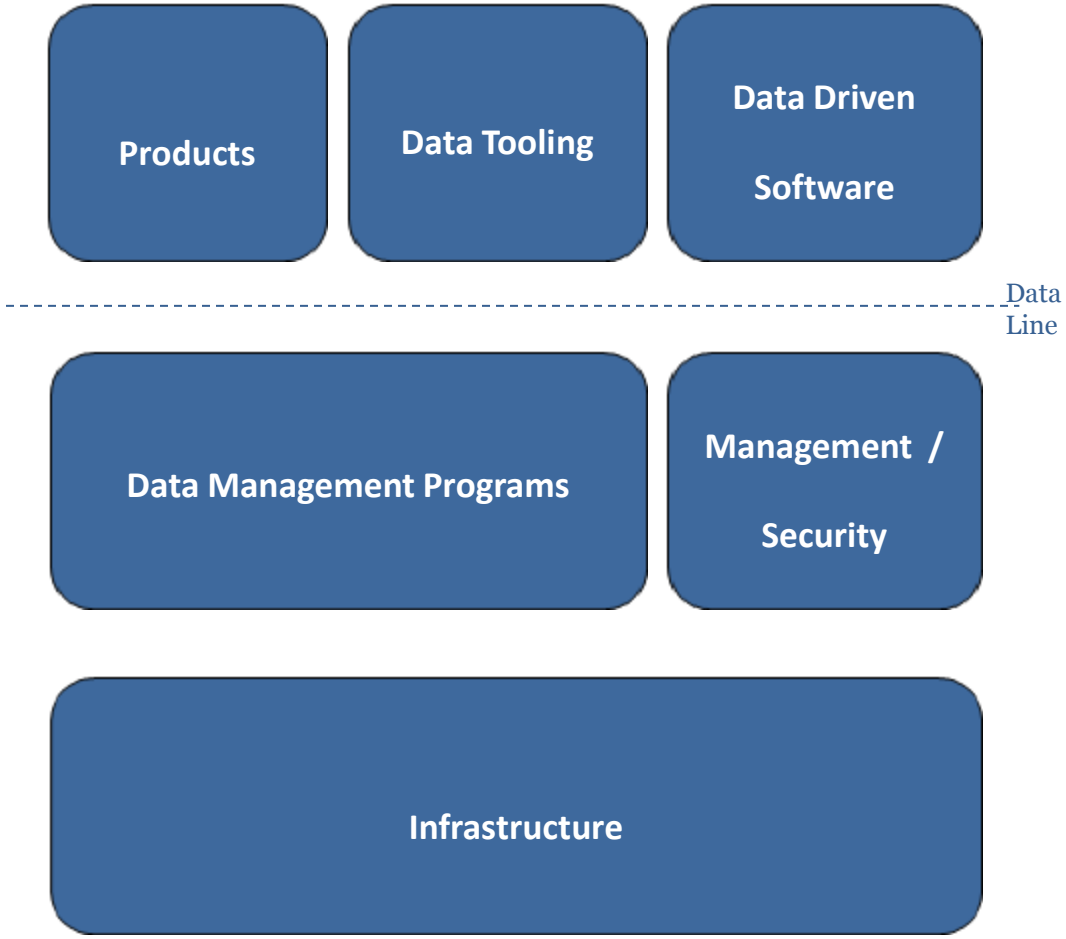
Consider Ecosystem Architecture -

Implementation and adaptation of new technology, as well as leveraging existing capabilities

- Hadoop (MapReduce)
  - Vendor selection
  - Integration with existing technology (hardware requirements)
- Leveraging existing tools and technology
  - Existing analytical tools (SAS, SPSS, R) will also still be useful with Big Data

Big Data Stack – Overview

The big picture of the Big Data Stack is reasonably simplistic



http://venturebeat.com/2014/02/25/a-key-investor-walks-you-through-the-big-data-technology-stack-video/

Key Features:

- **Infrastructure**
  Mainly Hardware with bits of Software
  E.g. Storage, Cloud, Virtual Systems, Networking
- **Data Management Programs**
  Enables most of the top of the stack features
  E.g. Relational/Non Relational DB, Hadoop
- **Management and Security**
  Necessary to make all the stack members to function smoothly
- **Products**
  Leveraging Data and packing in a consumable format
- **Data Tooling**
  Business Intelligence component
  E.g. SAS, Informatica, Advanced Machine Learning
- **Data Driven Software**
  Optimized processes after ES, unbounded solution customized to client needs

Enabling the Future With the Choices You Make Now

Do develop technologies and processes that are flexible enough to cope with whatever the future brings

Do not build something that cannot flex to future needs