

Big Data Analytics Project Management: Methodologies, Caveats and Considerations

Tiffani Crawford, PhD



Speaker and Contact Information

Tiffani Crawford, PhD

408-829-7096

zapquick@aol.com

Big Data Analytics Project Management book by Tiffani Crawford – See Amazon Books at <http://www.amazon.com> and search by title and author.

Tiffani Crawford, PhD, builds global Big Data Analytics systems. She has 20 years of high technology experience with Fortune 500 companies, including Cisco Systems, Cognizant, Bank of America, VISA/Inovant, BAE Systems, Applied Competitive Technologies, Ditech Networks/Nuance, Big 4 financial firms, defense contractors and startups. She has worked in seminal technology development in Big Data, analytics, cloud, networking, telecommunications, software development, distributed multi-tier applications, multimedia/digital, geographic information systems, intelligent transport systems, finance, security, policy systems and structural equation modeling. She is a credited software developer and published author. She earned her PhD from the University of Southern California in 2005. She has also earned her Master's, Bachelor's and various technology certifications. She is a member of PMI with various philanthropic contributions.



Introduction and Contents

Business Case for Big Data Analytics

- What is Big Data Analytics?
- Getting to Business Value
- Opportunities and implications
- Clients

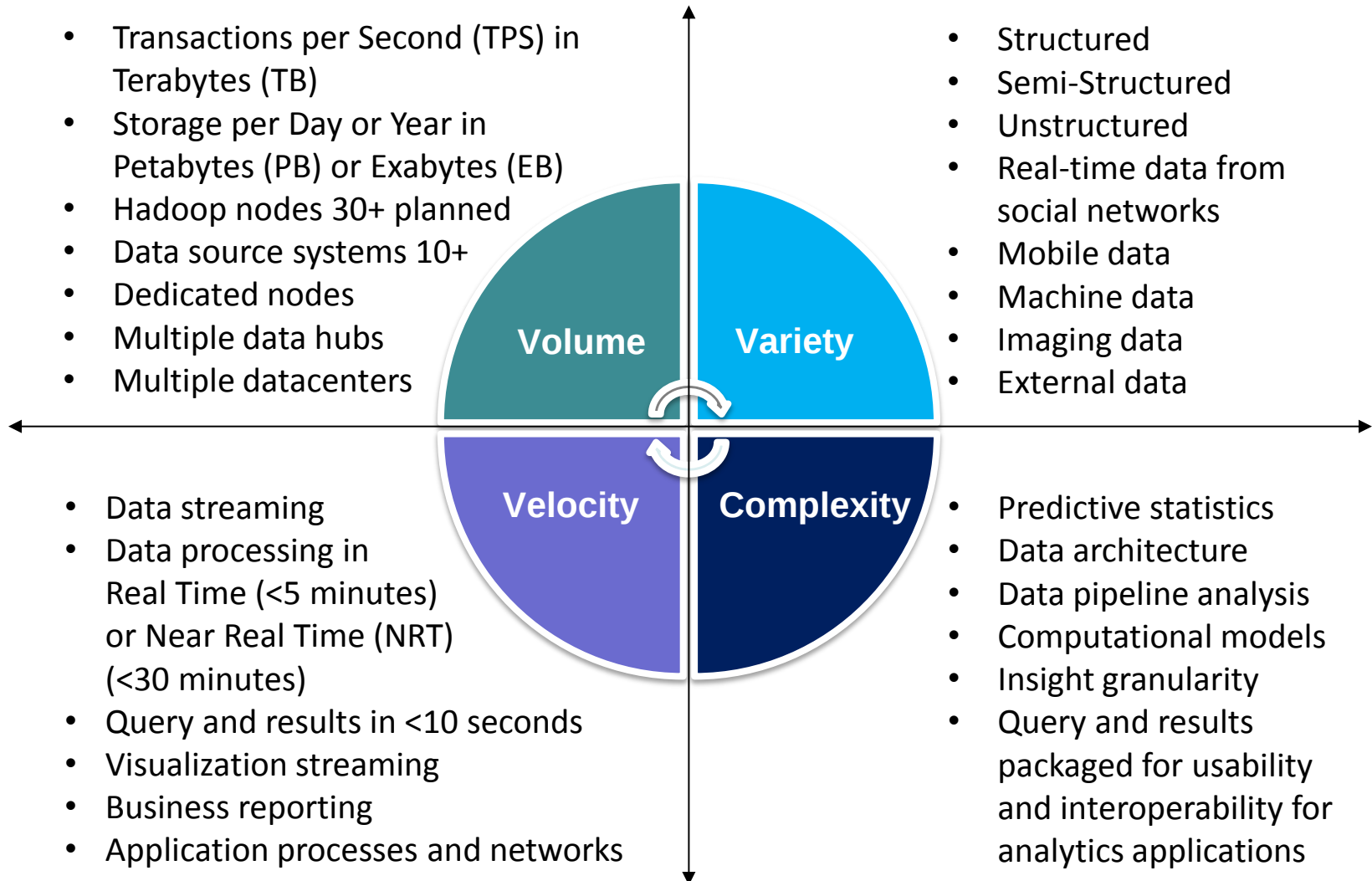
Project Management for Big Data Analytics

- Common misconceptions
- Does Big Data Analytics fit your client's interest, situation and experience?
- Choosing the best methodology
- Evaluating caveats and risks
- Unique considerations
- Growth and sustainability

Hands-On Exercise

- Create a repeatable model for a client's Big Data Analytics in a focus area that inspires you and your team
- Continue this exercise and discussion via social media and online applications

Big Data – INCOMING!!!!



Big Data Characteristics



Large quantities of many data types

- Structured
- Semi-Structured
- Unstructured
- Human
- Machine

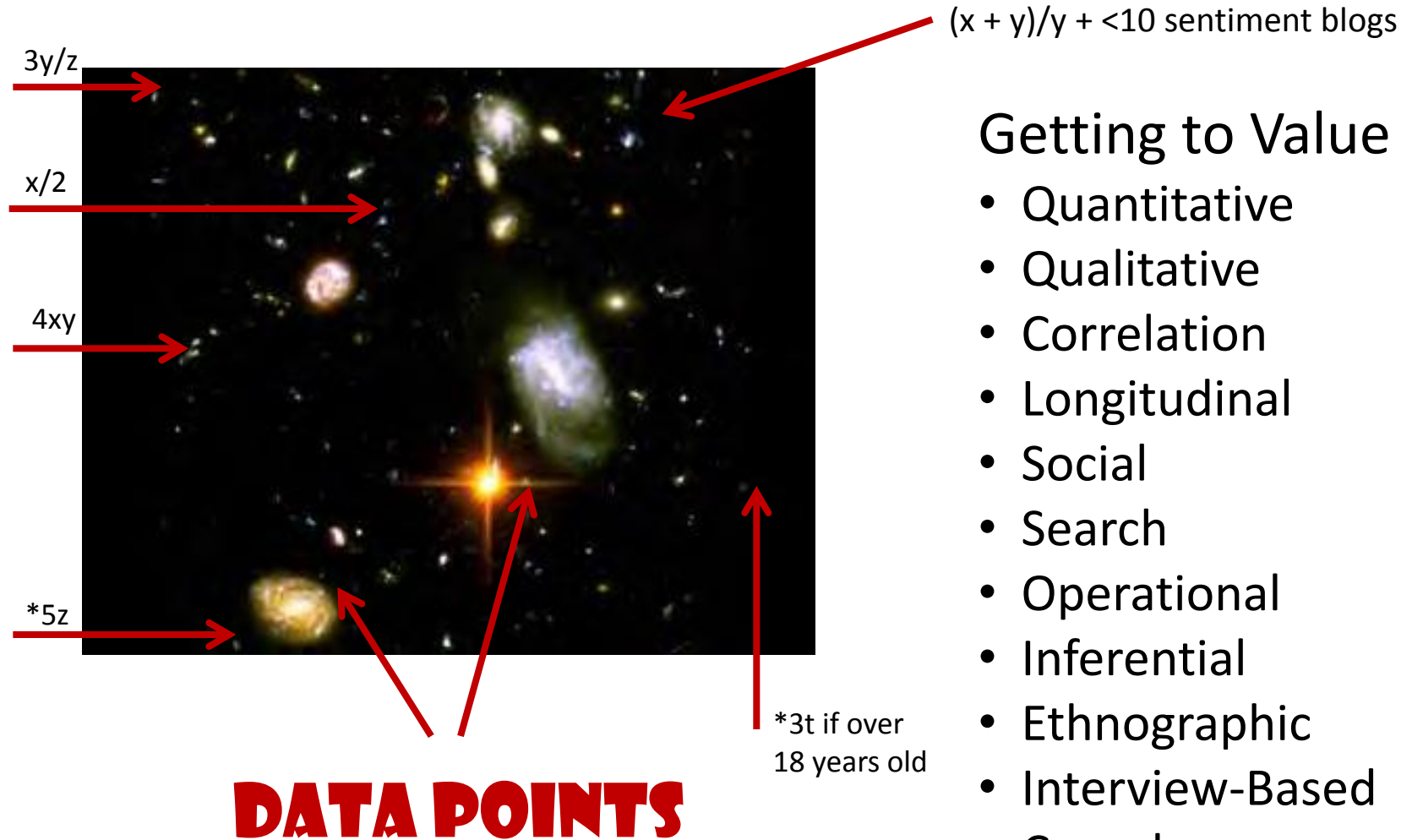
Valuable Insights



Getting to Value

- Quantitative
- Qualitative
- Correlation
- Longitudinal
- Social
- Search
- Operational
- Inferential
- Ethnographic
- Interview-Based
- Casual

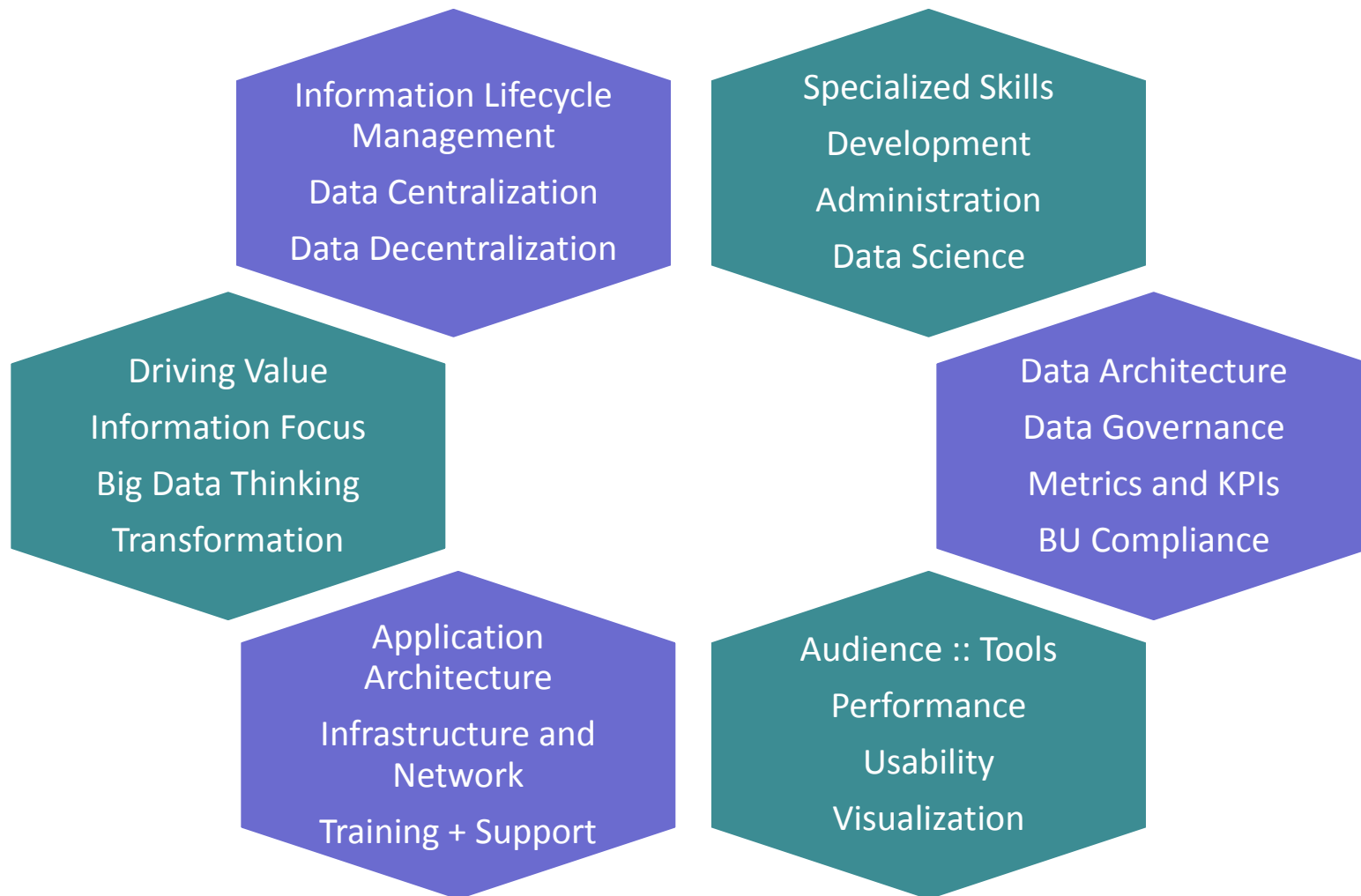
Valuable Insights



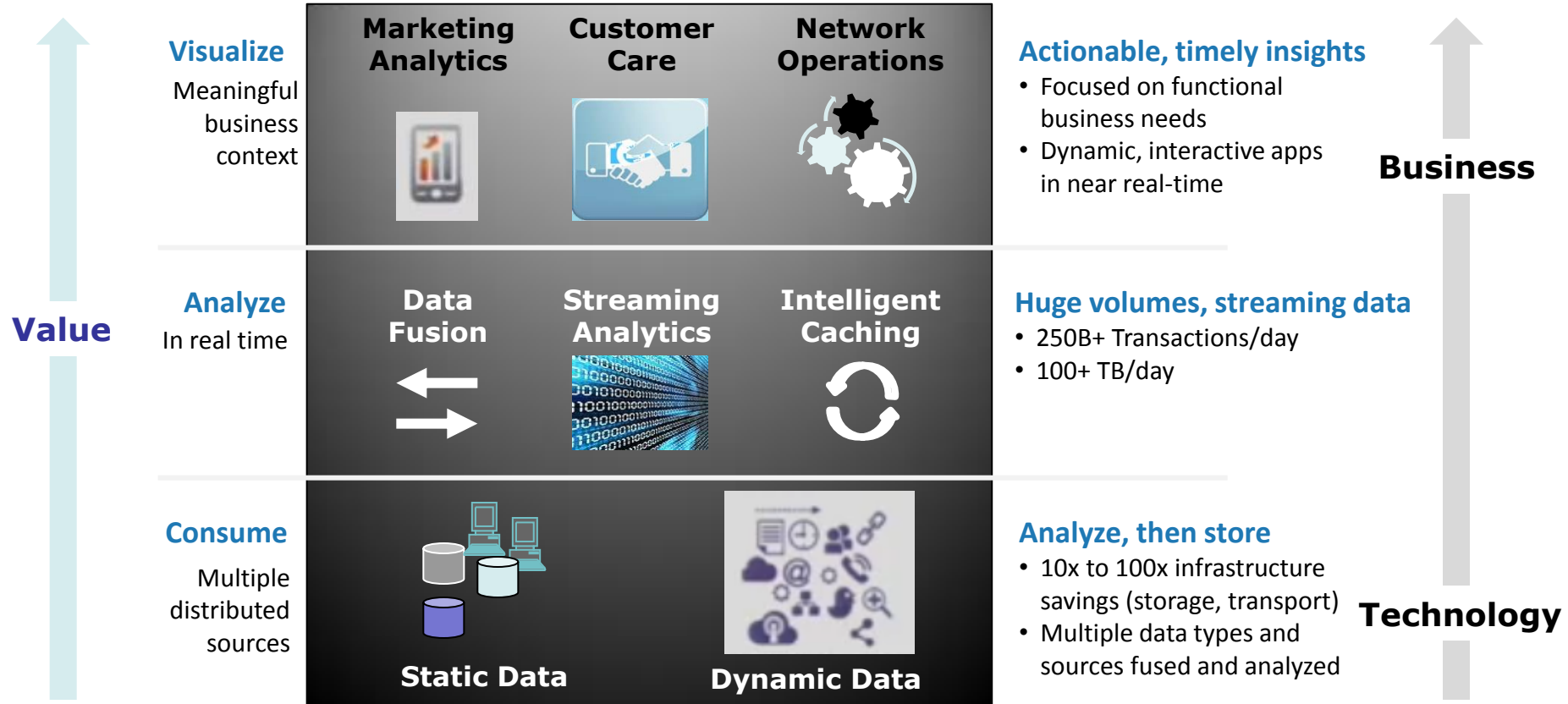
Getting to Value

- Quantitative
- Qualitative
- Correlation
- Longitudinal
- Social
- Search
- Operational
- Inferential
- Ethnographic
- Interview-Based
- Casual

Big Data Initiative Challenges



Technology to Business Value





Business Case for Big Data Analytics

- Applicable to all enterprise applications and clients
- Faster
 - High-speed data processing
 - 1 big job => many small jobs (reduce complexity)
 - Redundancies prevent data loss
 - Temporary storage is faster
 - Long-term storage is easier to access
 - High-speed query and results processing (<10 sec)
 - High-speed predictive analytics & responsive models
- Cheaper
 - Low cost (2-10x cheaper) up to massive scale
 - Technology-agnostic
 - Commodity hardware
 - In-Memory options
 - Interoperates easily with plug-ins
 - Open source available



Business Case for Big Data Analytics

- Easier
 - Data validation and quality = near-100% accuracy
 - Network resilience and 5-9s performance
 - Security resilience and management
 - Virtualization and automation
 - Network traffic management
 - Transitory data management
 - Data retention policies and management
 - Enterprise-wide analytics with quality data
 - Query management and reusable code
- Start Small and Practice
 - Define business value, use cases and metrics/KPIs
 - Open source, cloud or large supplier install base
 - Proof of Concept, Proof of Value, Proof of Technology
 - Early analytics modeling, Proof of Model
 - Pilot, testing, pre-production and production buildouts



Who uses Hadoop?



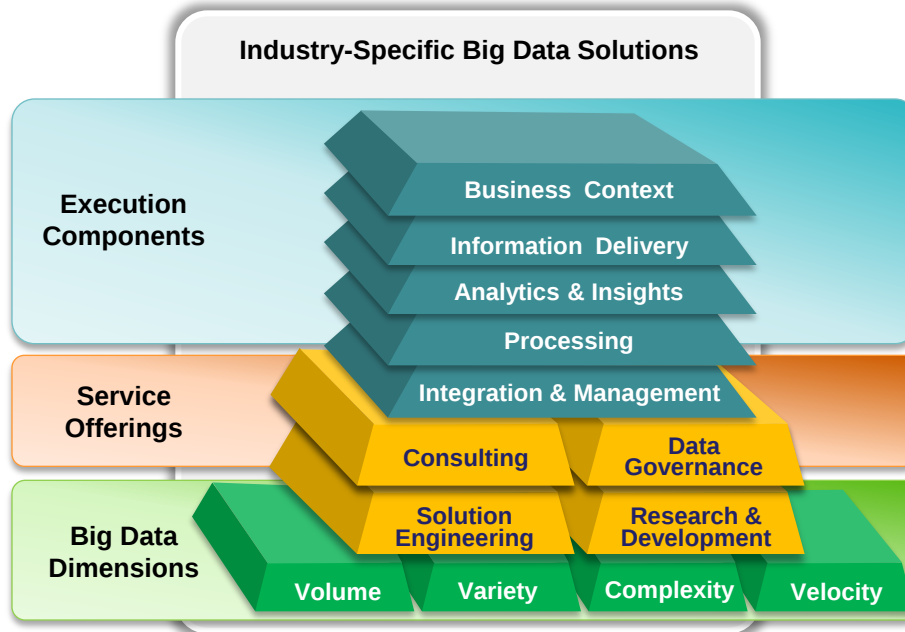
LTRC, IIIT-Hyderabad



The New York Times



Hadoop Technology-Agnostic Market



Research & Development
 Entrepreneurship
 Business Transformation
 Market Extension





4 Pillars – Complete Ecosystems

Hadoop

- Java or alternatives
- SQL
- Components in Ecosystems

NoSQL

- Alternative to SQL databases
- Components in Ecosystems

Hybrid Systems

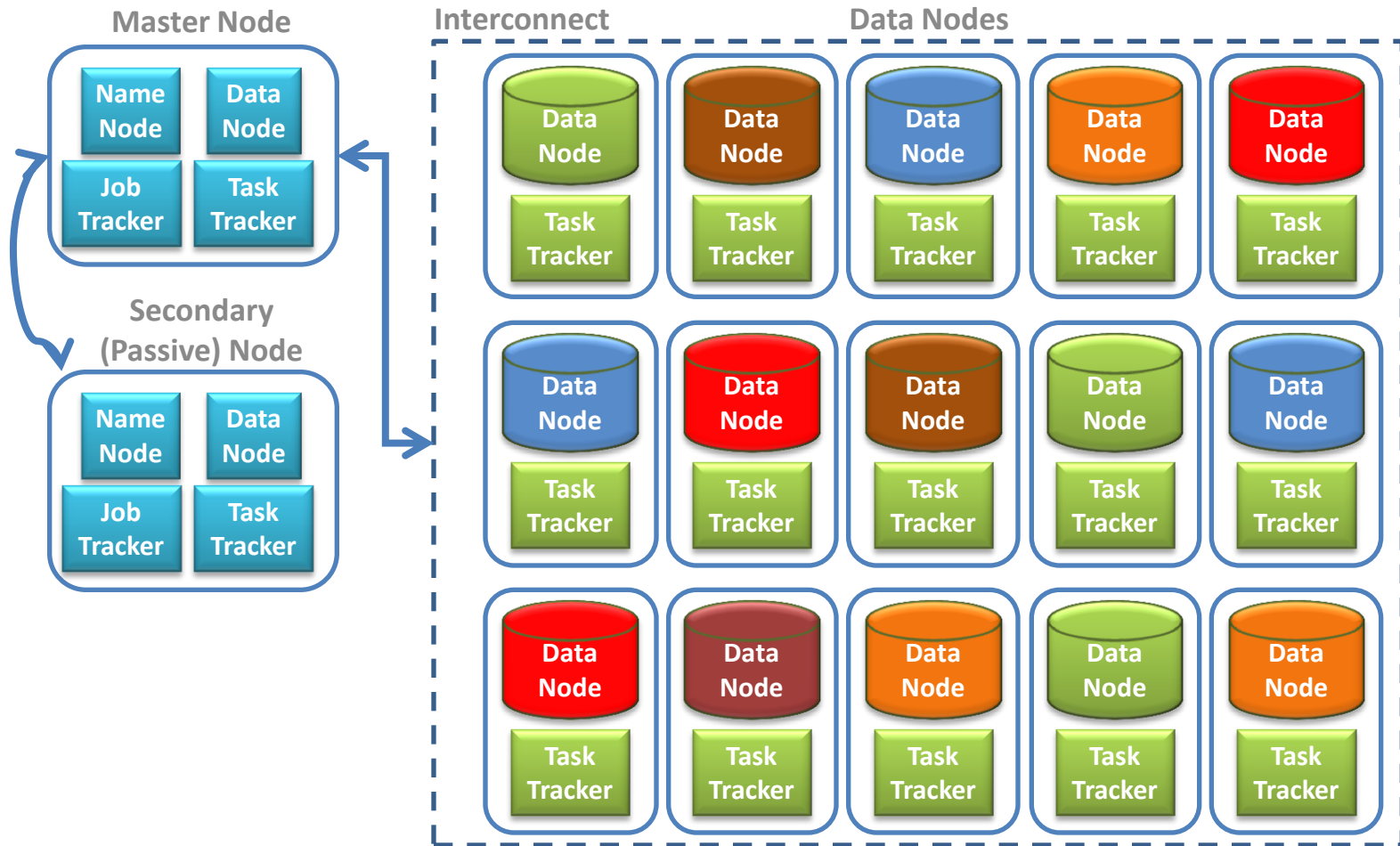
- Hadoop and/or NoSQL components
- Legacy databases and Enterprise Data Warehouses (EDWs)
- Legacy analytics for understanding what happened
- Install bases from large providers
- Slow ETLs moving data

Data Science

- Analytics for prediction and business transformation
- Data visualization and reporting



Hadoop = MapReduce + HDFS





Hadoop Ecosystem

MapReduce	Software framework for clustered, distributed data processing
HDFS	Primary storage system used by Hadoop in a distributed environment
Hive	Data warehouse
HBase	Data storage for distributed large tables for random, real-time read/write access to Big Data
Scribe	Log collection
ZooKeeper	Workflow management service
Avro	Data serialization
Chukwa	Data collection system to monitor distributed systems
Sqoop	Open source data integration tool to integrate data into HDFS from sources outside of Hadoop
Pig	High-Level Query Language Platform for analyzing huge data sets, which involves complex rules
Cassandra	Highly scalable NoSQL database which brings fully distributed design & BigTable's data model



NoSQL

- NOT relational algebra (RDBMS tables/relationships)
- SQL is not the data manipulation language





Hybrid Systems





Data Science

Making sense of data is both Art and Science

- ✓ Generate better insights
- ✓ Gain confidence in decisions
- ✓ Visualize the data
- ✓ Understand the data and communicate that
- ✓ Learn how to learn and adapt with agility

Methods

- ✓ Data mining
- ✓ Machine learning
- ✓ Artificial intelligence
- ✓ Information retrieval
- ✓ Statistical analysis
- ✓ Gap analysis



Analytics Programming Languages

Top 10

- **R**
- **Python**
- **SQL**
- **SAS**
- **Java**
- **MATLAB**
- **High-level data mining suite**
- **UNIX shell/awk/sed**
- **C/C++**
- **Pig Latin, Hive and other Hadoop-based languages**

Analytics and Data Visualization

syncsort

Business Objects



Datameer

Google BigQuery

Spotfire®
TIBCO Software

JASPERSOFT

sas

pentaho

NETEZZA

COGNOS®

+ a b l e a u

CAMBRIDGE
SEMANTICS

MicroStrategy®

DOMO

Microsoft

informatica

ORACLE®

semantelli



Information
Builders

QlikView

CLARABRIDGE

IBM SPSS



Data Visualization

Context

Start Date

NOV
MON 05



End Date

NOV
MON 05



Competitors

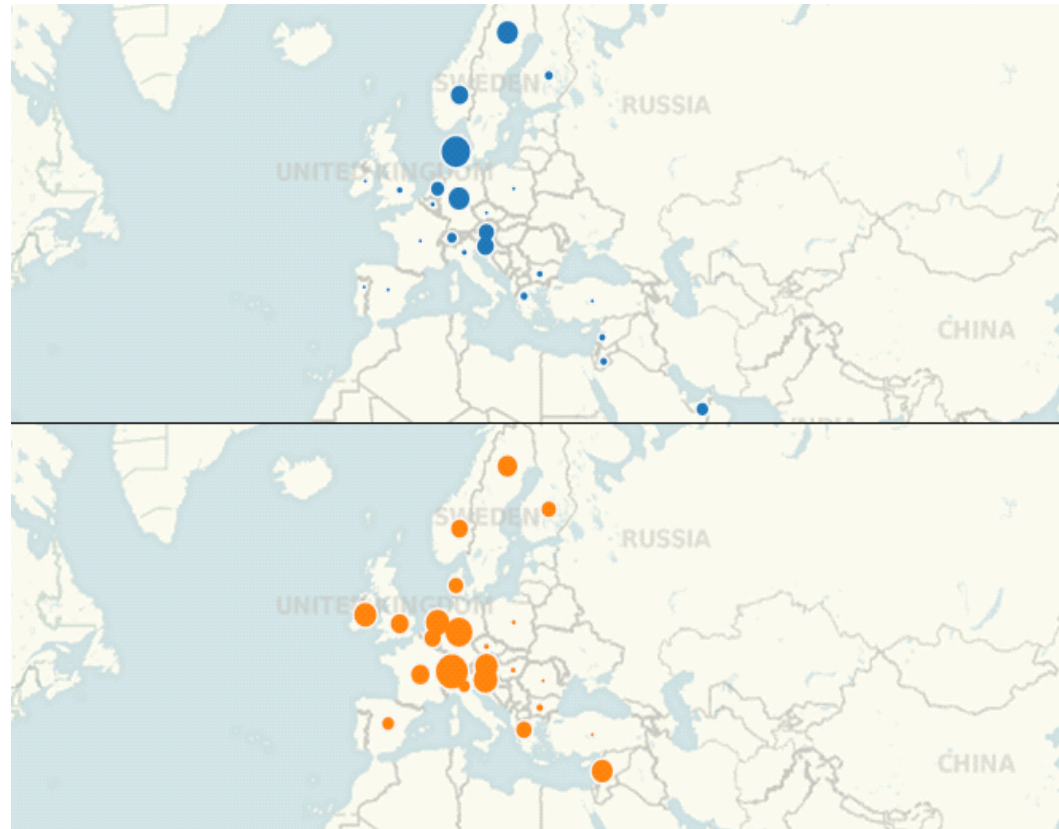
- ☐ Competitor1
- ☐ Competitor2
- ☐ Competitor3
- ☐ Competitor4
- ☐ Competitor5

Sentiment Heat Map

Context: Sample

Start Date: 01-Jan-2012

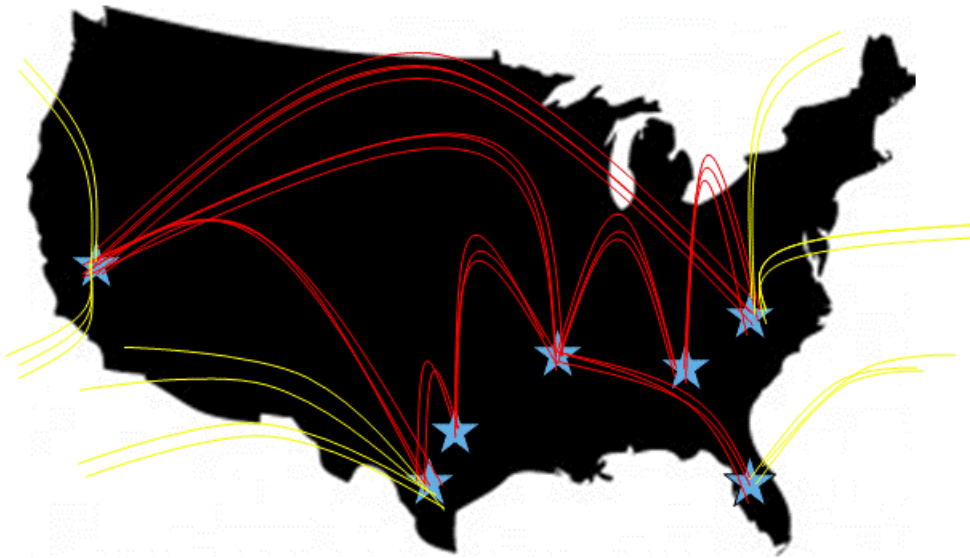
End Date: 31-Dec-2012





Data Visualization

Capacity Management
(Interstate Traffic – Core)



= Content Provider Network Sites

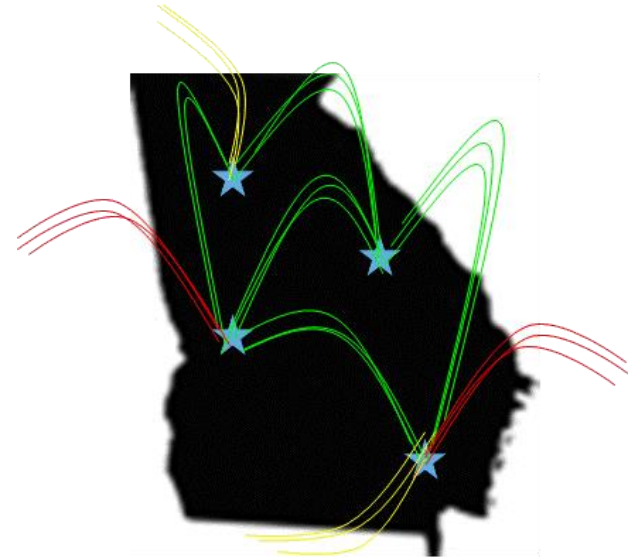


= Interstate Traffic (CORE)



= Interconnections (ENNI, NNI, Colo, etc.) – Leased Facilities

Capacity Management
(Intrastate Traffic – Transport)



= Content Provider Network Sites



= Intrastate Traffic (Transport)



= Interconnections (ENNI, NNI, Colo, etc.) – Leased Facilities



= Interstate Traffic (CORE)

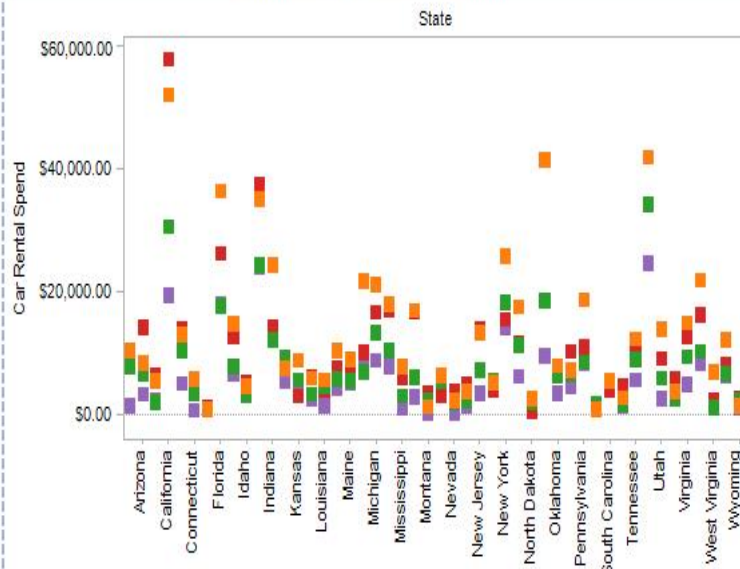
Data Visualization

Link Analysis Dashboard

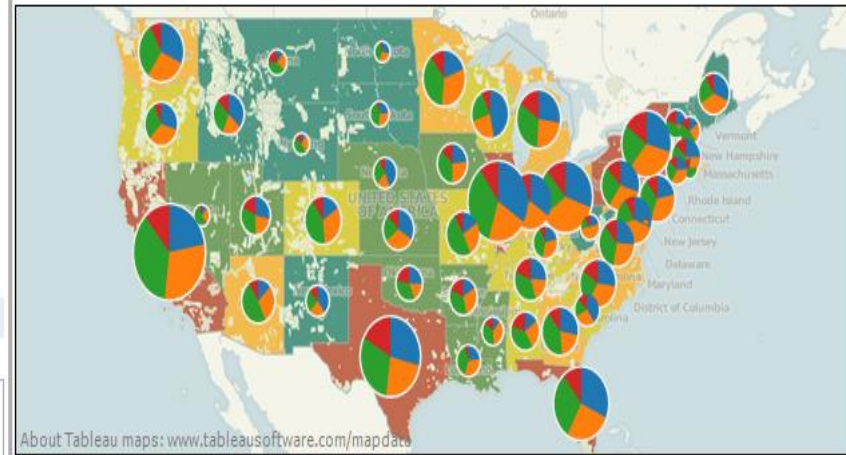
Customer Link Analysis For a Credit Card Company (Using Travel Merchant Data and Car Rental Merchant Data)



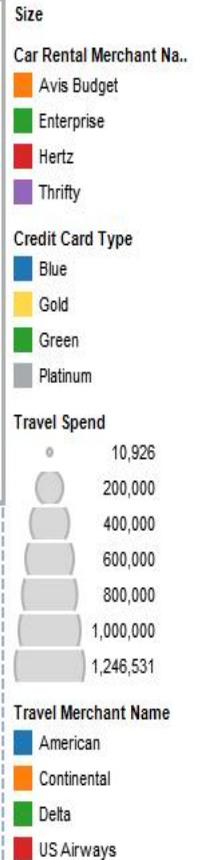
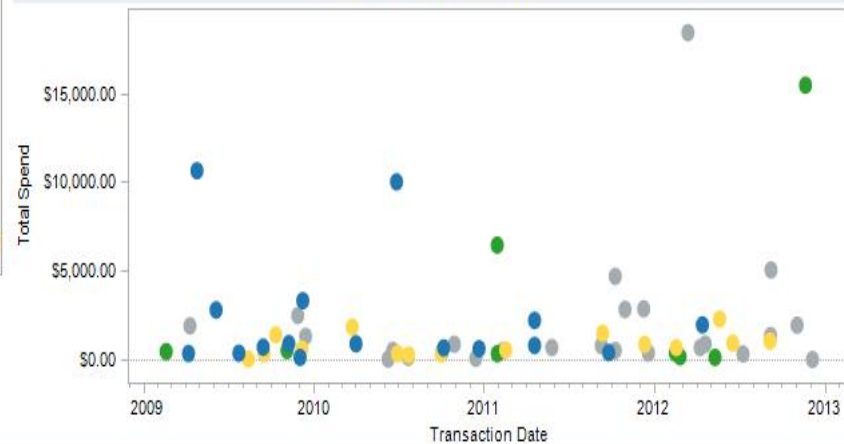
State wise Car Rental Pattern



State wise Travel Pattern



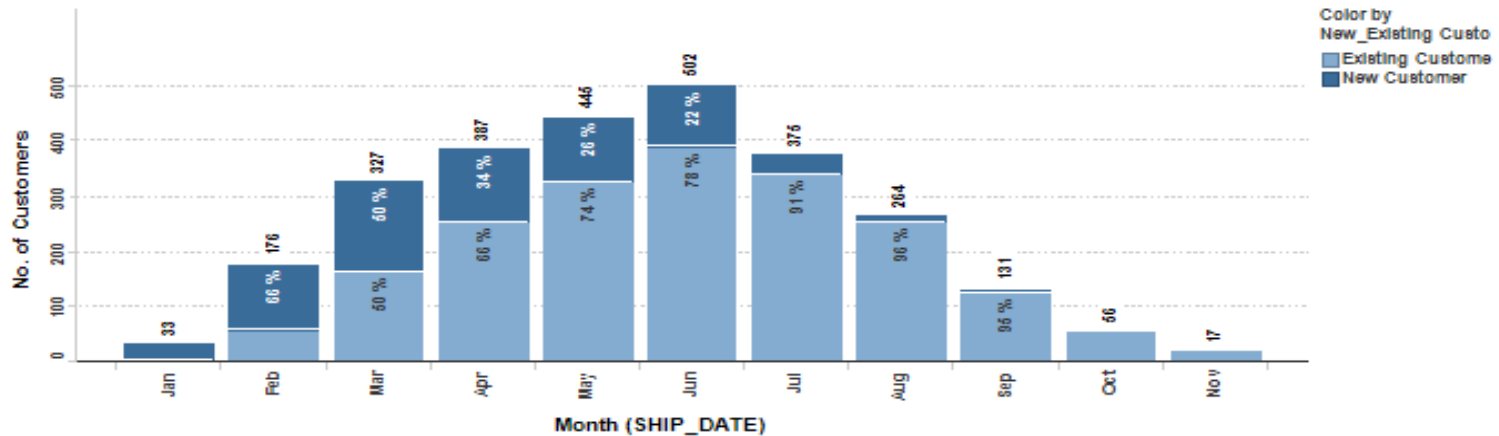
Customer Spend Analysis



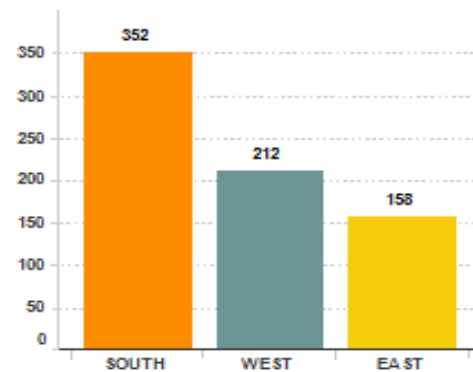
Data Visualization

Impact Now Customers

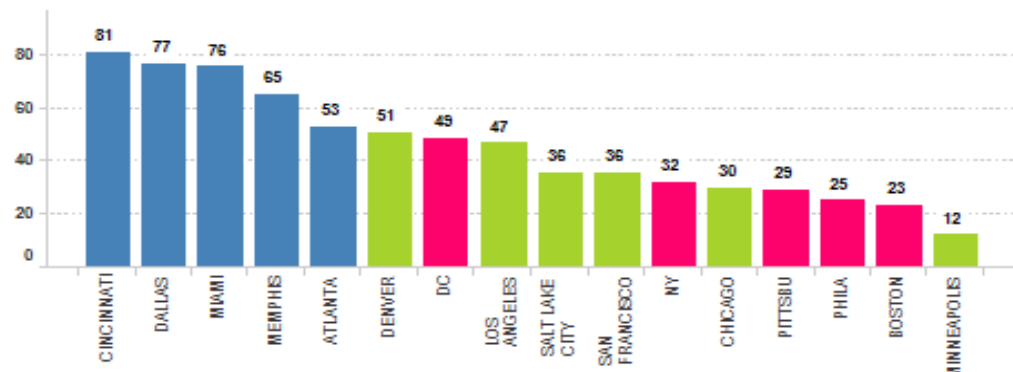
New vs Existing - No. of Customers



No. of Customers by Region



No. of Customers by Division

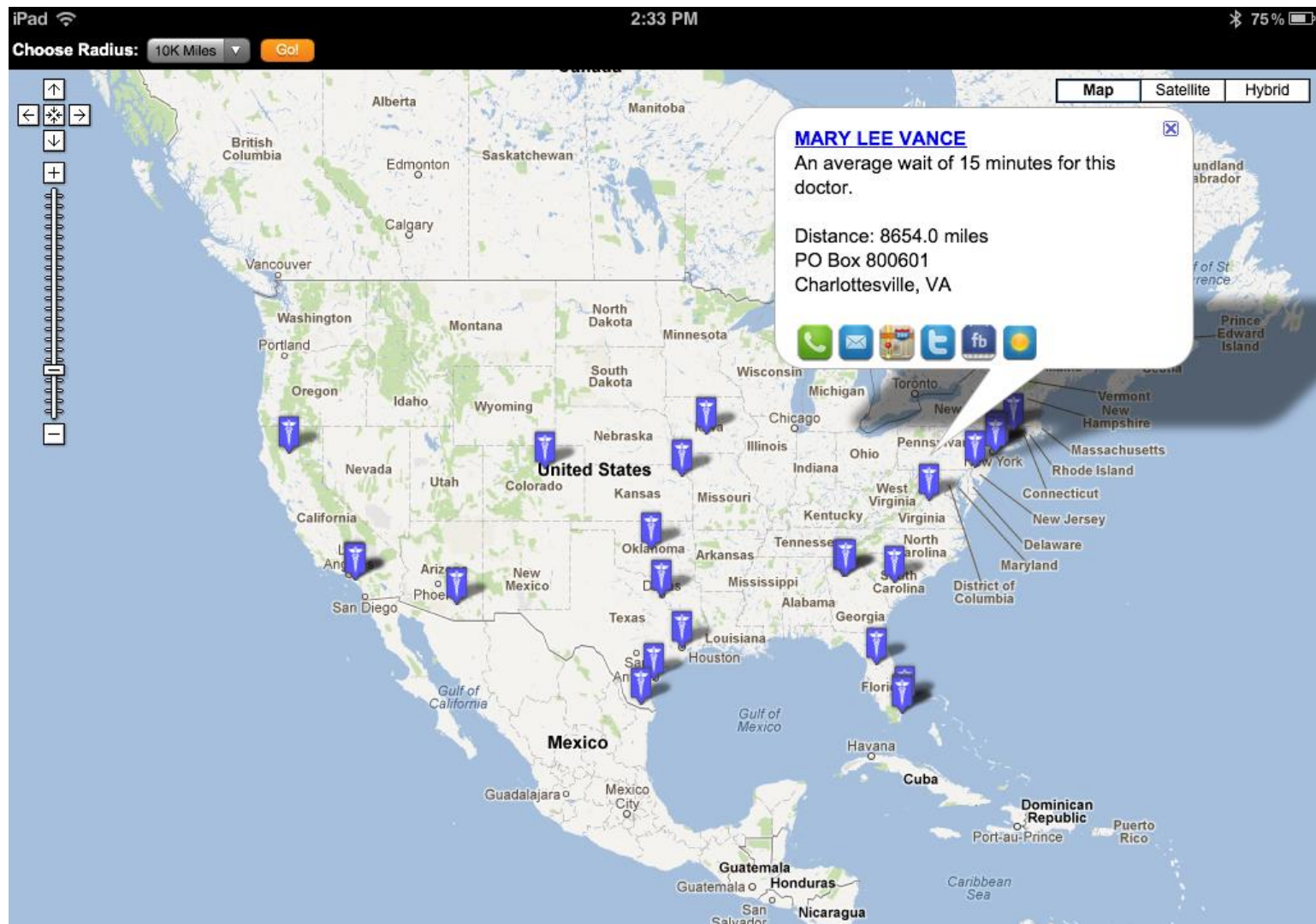


*Detailed data is available in Data Summary Report

Confidential Document - for Internal, Non-Promotional Use Only

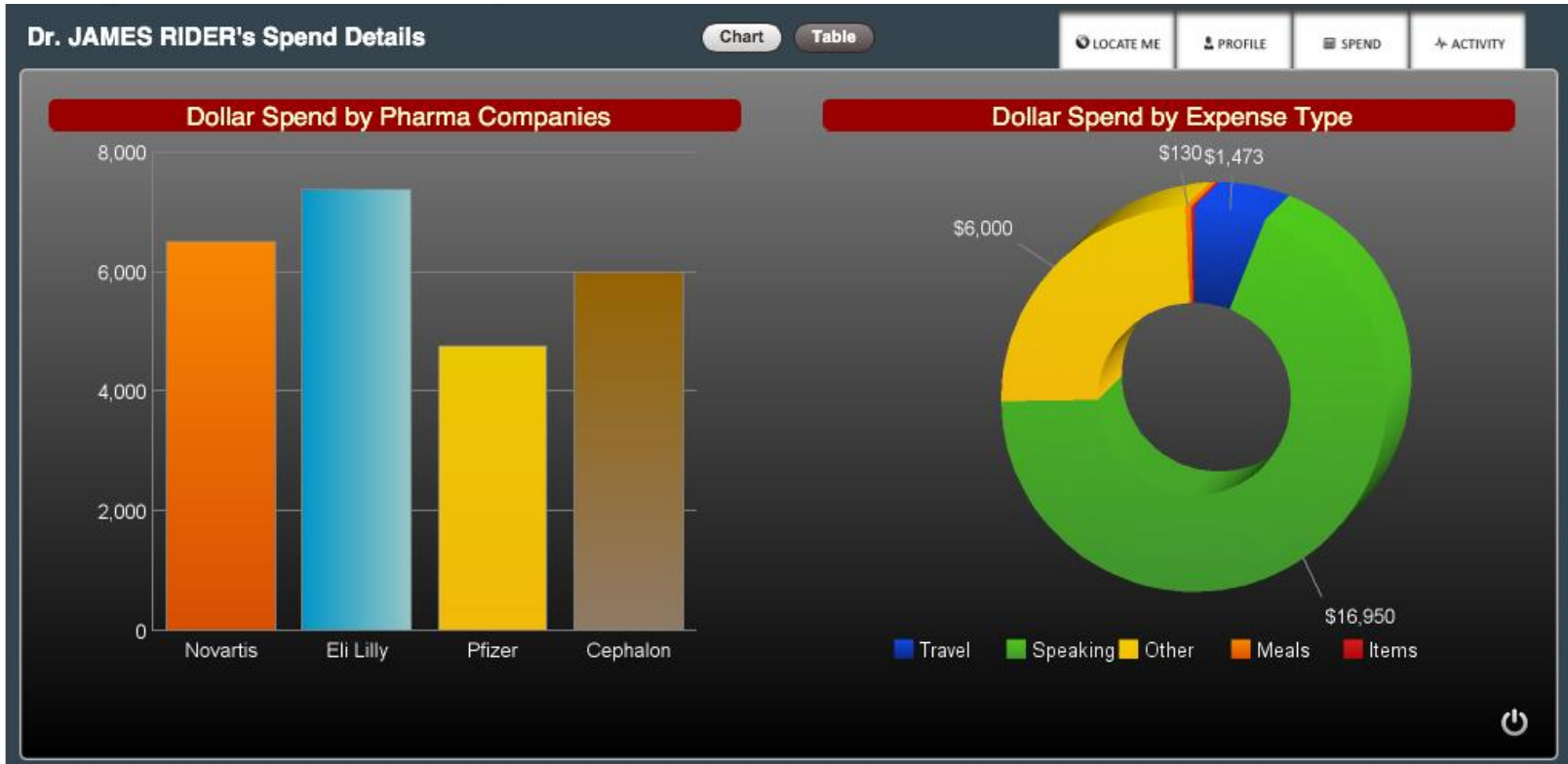
5

Data Visualization

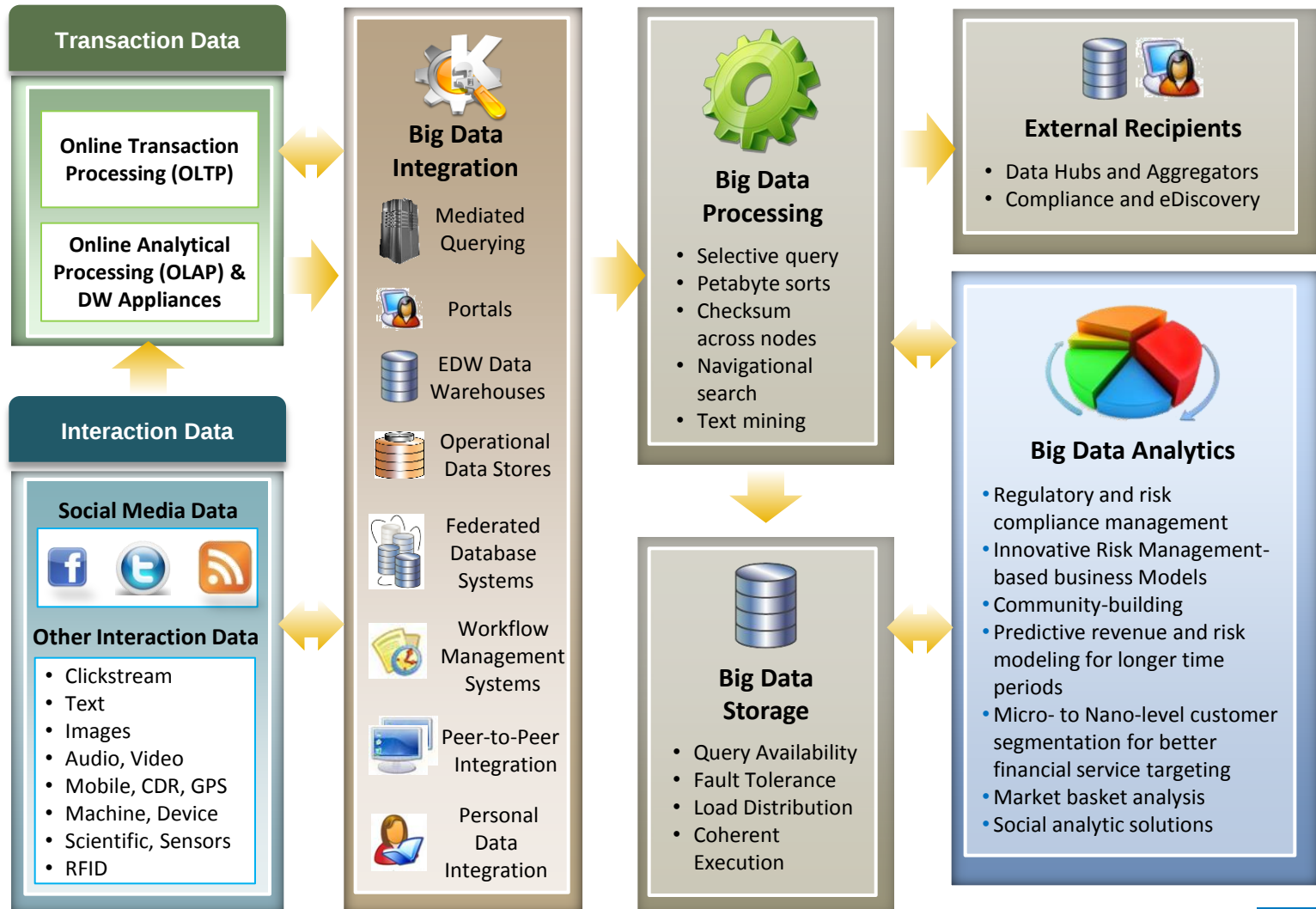




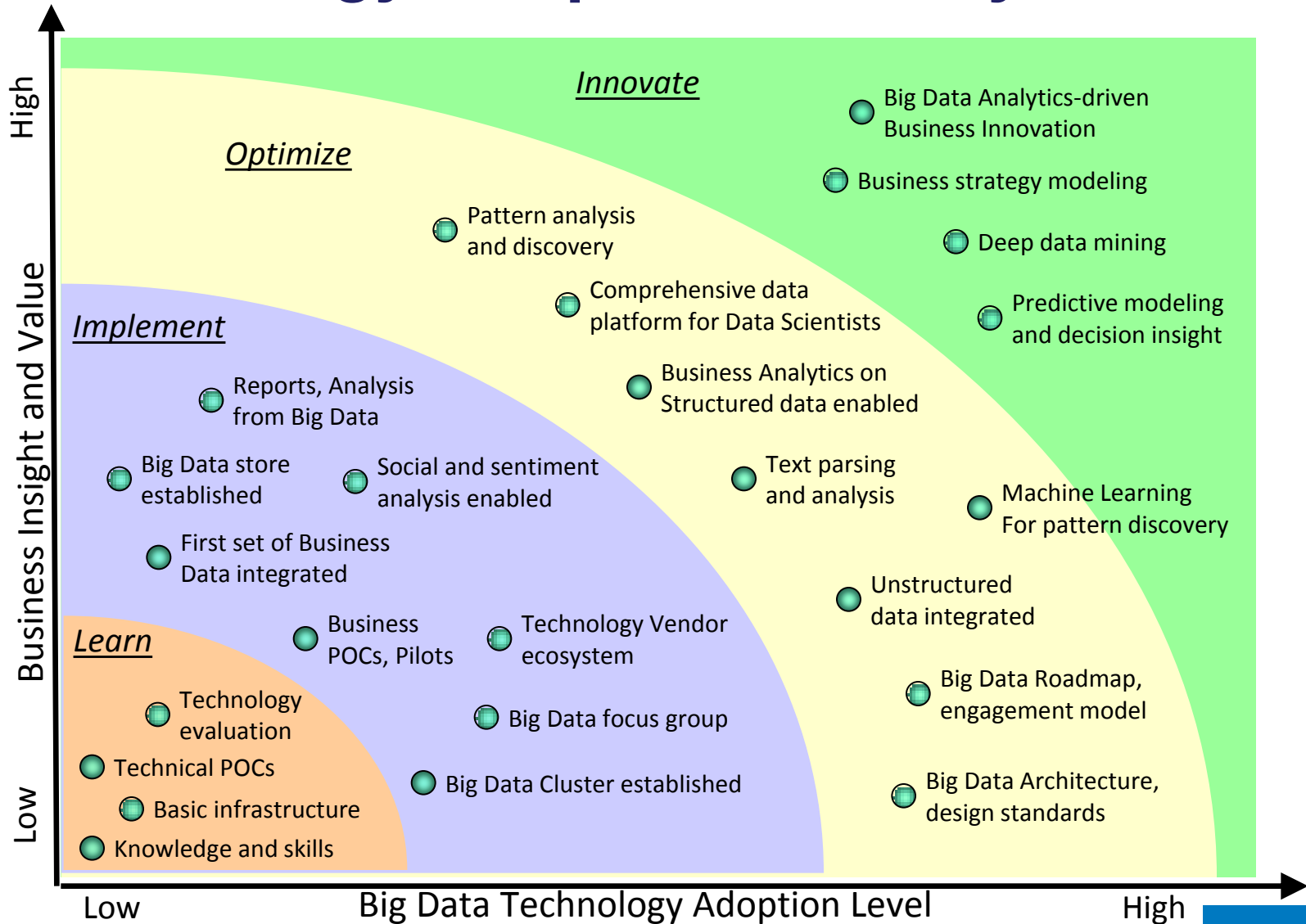
Data Visualization



Ecosystem Architecture – Think Layers



Technology Adoption Maturity



Forrester Wave – Enterprise Hadoop



Gartner Magic Quadrant – DW + Hadoop



TERADATA
Raising Intelligence



ORACLE®



Microsoft

EMC²



Gartner Magic Quadrant – BI + Hadoop





Opportunities and Implications

Business value delivery

- Research process automation, analysis and discovery
- Intellectual property
- Global, always-on data availability
- Reduced cost and rework
- Collaborative leadership
- Learning opportunities
- Rewards and fairness (vertical and horizontal equity)



Clients

- Enterprise data management, architecture, governance and validation
- Financial data management
- Operational data management
- Investment data validation
- Business development analytics and decision-making
- Data query-results systems with fast processing power
- 360-degree customer views
- Social media, people-interest matching and innovation discussions



Trends and Future

- Innovate
- Adapt
- Streamline
- Automate
- Consolidate
- Discover
- Extend
- Renew



Project Management

Where does Project Management fit in?

- Hadoop was used first in 2004 and became better known by 2007. It is 6 years old.
- Originally, Hadoop was thought to not need project management because it was so easy to get it together, up and running.
- Today, some clients process Petabytes of data per day globally. Transactions reach Terabytes per second.
- The definition of Big Data has expanded to include more technologies.
- Large infrastructure providers offer complex, expensive systems for vertical clients.
- Implementations can involve 50-100 engineers.



Project Management

Big Data Analytics Project Managers are required.

- Project Managers with technical expertise and business sense are needed.
- Expert judgment in working with the technologies is needed.
- Program Managers must help scale and synergize the efforts across projects.

Experts are hard to find in the job market.

- Hadoop developers and administrators – 1 expert for every 2 open positions
- Data scientists – 1 expert for every 4 open positions
- Hadoop Project and Program Managers – 1 expert for every 4 open positions



Project Management

Various sources report that 65-100% of Big Data Analytics projects fail.

- Incomplete
- Out of time
- Over budget

Sources reporting less failure than 65% include POCs, POTs, POVs and pilots in this estimate.



Project Management

What happened?

- Could not hire required staff early enough or at all
- Did not anticipate current data needs and new data handling
- Did not identify sufficient components and interfaces during design
- Did not scale some of the application components to the build phase
- Did not discover critical path needs across multiple teams
- Did not load balance and test performance during development, CI and ST
- Did not have clear data architecture and governance to validate data
- Took too long to approve QA test accounts for credit cards, mobiles, CSP traffic
- Network virtualization requirements were not clear
- Did not have time to document for support and enhancement work
- Did not have sufficient sponsorship and buy-in from client team or stakeholders
- Budget was pulled before results could be proven
- Phase 1 infrastructure was not purchased or installed in the datacenter
- Analytics could not be implemented



Program, Portfolio & PMO Management

External Impacts

- Big Data Analytics and Market Hype
 - Lots of high-level information available
 - Need for expertise
- Stakeholder Caution
 - Need to prove methodologies, prioritization criteria and results/ROI
 - Need to position for growth
 - Growth because we get it
 - Growth against competition's growth
 - Growth against internal resource competition and politics
 - Balance centralization and decentralization of data standards and access
- Vendor Selection
 - Many attempting players with limited experience, expertise and staff
 - New product lines that pair hardware and cloud software
 - Marketing-oriented, low-substance presentations to clients
 - Expansion of stakeholders from IT groups to Business and Marketing groups



Program, Portfolio & PMO Management

Internal Impacts

- Lack of predictive analytics about fast-moving Big Data Analytics projects
- Lack of clarity about project synergies of expertise and reusable components
- Lack of knowledge in selecting methodologies and coaching PMs



Program, Portfolio & PMO Management

Structuring the PMO Services and Portfolios, Programs and Projects

- Managing Requests
 - Questions and consultations
 - Help with clients, conferences and capabilities
 - Use cases, POCs, POTs, POVs and pilots
 - Staffing, training and coaching
 - Technology, strategy and business value
 - SLAs with clients and internal business units
- Managing the project hierarchy
- Managing methodologies
- Managing documentation



Program, Portfolio & PMO Management

Organizational Adaptation – Balance of leadership and environment

- Population Ecology – Survival of the fittest, with or without leadership
- Life Cycles – Each cycle needs Project Management
 - Creativity and entrepreneurship
 - Collectivity and sophistication
 - Formalization, control and efficiency
 - Elaboration of structure to decentralization and expansion
- Strategic Choice – Information, locations and moments of choice produce the actions of Prospectors, Analyzers, Defenders and Reactors
- Symbolic Action – Defined social construction of reality and roles



Program, Portfolio & PMO Management

Organizational Leadership

- Organizational Transformation
 - Coherence with general theories
 - Alignment with organizational frameworks and understanding of the work
 - Types of processes and ways of performing the work
- Reflective Leadership and constant communication
- Leadership as a collective, interactive practice
 - Leadership in different contexts by different people
 - Taking the perspective of the other people
 - Leadership is a relationship
 - Not a transaction
 - Not a relational context such as a hierarchy or functional role
- Transformational Leadership
- Organizational Learning
- Ambidextrous Organizations



Program, Portfolio & PMO Management

Thriving Projects and Project Managers

- Understand and help the client
- Collaborate with the client, elicit needs and elaborate innovations
- Deliver a rewarding scope of work and business value
- Enrich the team with good relationships and interesting, rewarding work
- Remember work-life balance for everyone



PMBOK and Big Data Analytics

Impacted areas that require Big Data Analytics expertise

- Less delimitation of Project, Program and Portfolio Management
- Project Lifecycle
- Stakeholders and Organizational Influences
- Organizational Process Assets
- Project Management Processes
- Project Management Wisdom

Do we need a special methodology, BDA-PM?

- We need specific expertise and general understanding among stakeholders.
- We need formalized best practices.



Typical Project Lifecycle

The lifecycle can be bumpy.

- Highly iterative
- Highly exploratory
- Team is involved longer due to the need for more testing and support

This can lead to the impression of:

- Lack of awareness and control
- Lack of communication
- Lack of planning

Set client, internal stakeholder and project team expectations for flexibility.



Typical Project Lifecycle

Initiate, Plan, Execute and Close in short phases for a series of smaller cycles

- Use models, POCs and pilots to minimize risk and explore options easily
- Work in parallel, overlap phases and exploit efficiencies of scale
 - Build for re-use and redeployment
 - Code bundling and coding maturity
 - Environments (pilot, QA, staging, pre-production, production)
- Design for testability and define testing needs early to avoid delays of a few weeks to months depending on the internal and external services needed
- Design for easy application and infrastructure monitoring

Governance is needed. Escalations may not solve problems.

- Don't run so fast that only critically-escalated work is completed
- Collaborate to solve problems but don't design by negotiation
- Know when the team needs facilitation or simply sleep



Organizational Process Assets

Needed OPAs do not exist in a complete, clear form

- Many traditional EDW and BI assumptions are no longer valid
- Templates may be inapplicable or need adaptation

Estimations of system volumes, capacity and load balancing require expertise because Big Data Analytics systems process data differently

New technology brings new knowledge

- Knowledge transfer
- Training and coaching
- Documentation of design, implementation and support
- Knowledge base development
- Coding standards
- Best practices
- Code libraries
- Reusable frameworks, tools and scripts



Organizational Process Assets

Privacy and security constraints

- Delimitation of tasks, access and visibility
 - More coding for data selection and masking
 - Encrypted databases
 - Security testing of internal code and externally facing systems
- Permission requests with formal approvals to view assets
- In-code documentation prohibited



Project Management Processes

Types of Processes

- CMMI is very helpful!
- Processes must be lightweight and highly facilitative
- Traditional process weight may be too heavy for Big Data Analytics iterations

Big Data Analytics must be built as a complete ecosystem in order to function well

- Use the roadmap, project plan and high-level WBS to identify missing pieces
- Manage with milestones and progress toward them

Estimates have some value but may be off considerably

- Waiting for estimates does not tend to make them more accurate
- If abstraction is needed, use modified Fibonacci for Agile software projects



Schedules

Scope flexibility may be limited because a complete application must be built

- Regulatory compliance must be in place
- Application extensions are important, including Search and other services
- Automated testing must be coded and used
- Performance testing is required
- Monitoring and network automation are required
- Backup and failover testing is required
- One security incident can sink the product
- Users must like the product

Schedule flexibility may be limited

- Infrastructure acquisition, shipping, installation and verification
- External testing services and other new vendor contracts



Quality

Quality is easier, but expertise is required to identify requirements

- Data validation
- Data reconciliation and deduplication
- All-or-nothing transactions
- Data privacy
- ACID and BASE
- Dedicated nodes



Staffing

Staffing Shortages

- Company must be willing to pay more for qualified people
- Delimit tasks on an expertise basis and push lower-level tasks to learners
- Build people to have more expertise
- Hire people with transferable skills and train them for more



Communication

Communication must be collaborative and interactive

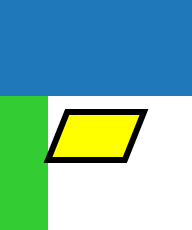
- Define interfaces
- Do not over-ask for estimates and status updates
- Do not assume that escalation occurred appropriately
- Problem-solving may follow a path of rapid tribal knowledge
- Expertise produces efficiencies
- Most problems are real and need solutions



Risk Management

Risk Management and Mitigation

- Substantial expertise in Big Data Analytics is required
- Risks can be estimated and mitigated
- Be more specific than the traditional top 5 risks



Common Misconceptions

- “Big” does not mean that it’s a monolith project. Start small and grow. Capacity management planning is easier.
- Waterfall is not required simply because it’s “Big”. Your client who loves waterfall can still see regular reporting without your development teams suffering through the methodology, risks and other implications.
- Few clients are actually “Dinosaurs”. Really.
- Hype does not mean that more burn-in is needed. Hadoop has been in use in enterprises since 2005, picked up steam in 2007 and expanded exponentially each year since then.
- Big Data processing or storage alone does not bring the business value.
- Analytics without data architecture, governance, validation and quality do not bring the business value.

Overcoming Obstacles

Typical obstacles that inhibit 'Next Best Action' or other multi-channel marketing capabilities range from organizational to analytical

Approach

Incremental steps reduce gaps

- Subset of channels, programs and members
- Process design + learning system
- Framework for scaling

DESIRED STATE

Real-Time Triggers Submitted



Analytically-driven Real-time offer engine



Next Best Action

CHANGE MANAGEMENT

- Organizational discipline to embed analytically-driven process into unified framework for offer management

ANALYTICAL

- Lack of channel preference segmentation, propensity scores by channel and program
- Absence of learning system for refinement

DATA INTEGRATION

- Data resides in silos
- Lack of 360-degree view of customer

ORGANIZATIONAL

- Programs "owned" by different groups and/or third parties
- Different priorities and varying approaches

CURRENT STATE

Data Submitted

No Action





Does Big Data Analytics fit your client?

Level of Analytics

- Client is Crawling (Storage only)
- Client is Walking (Basic Analytics for reporting what happened)
- Client is Running (Predictive Analytics for forecasting and proaction)

Data Management

- Client has no formal data governance
- Client has centralized or decentralized data management



Does Big Data Analytics fit your client?

Environment

- Big Data Analytics is only part of the client's environment
- Client wants the cloud
- Client requires on-premise technology
- Client uses only SaaS providers and has no EDW or Data Science

Readiness for Big Data Analytics

- Client has no metrics for assessing business value
- Client has no transformational change management best practices
- Client has no in-house technical or other expertise



Choosing the Best Methodology

Agile

- **Scrum works best because of the iterative discovery process**
- RAD
- Customer-Driven
- Stanford Advanced Project Management
- Others

Waterfall and Hybrids

- Test-integrated
- Requirements in advance



Scrum and Scrum Again

Teams

- Use several scrum teams
 - Development
 - Unit and integration testing
 - QA, automation and performance testing
 - Infrastructure buildout
 - Application scripting
 - Application configuration of workflows, dictionary, etc.
 - Network virtualization and monitoring
 - Support
- Leverage knowledge as it is built



Scrum and Scrum Again

Coordination

- Have a daily Scrum of Scrums and report results to management
- For 24x7 global teams, have a handoff scrum
- For the same team running double shifts, scrum morning and evening
- For infrastructure buildout, scrum morning and evening
- Audit completions daily
- Do not ask for additional updates!



Evaluating Caveats and Risks

Start with a small environment

- Do a Proof of Concept, Proof of Value or Pilot
- Use cloud services (AWS, Cloudera, etc.)
- Save money by building out the Pilot to Pre-Production and Production

Start with a well-defined core dataset

- What data is available?
- What analytics questions are the most interesting?
- How will we add new data types?
- How will we add new analytics?



Evaluating Caveats and Risks

Define well, but you do not need to know everything in advance

- Processes and workflows
- External integration
- Component decomposition and messaging
- Necessary services

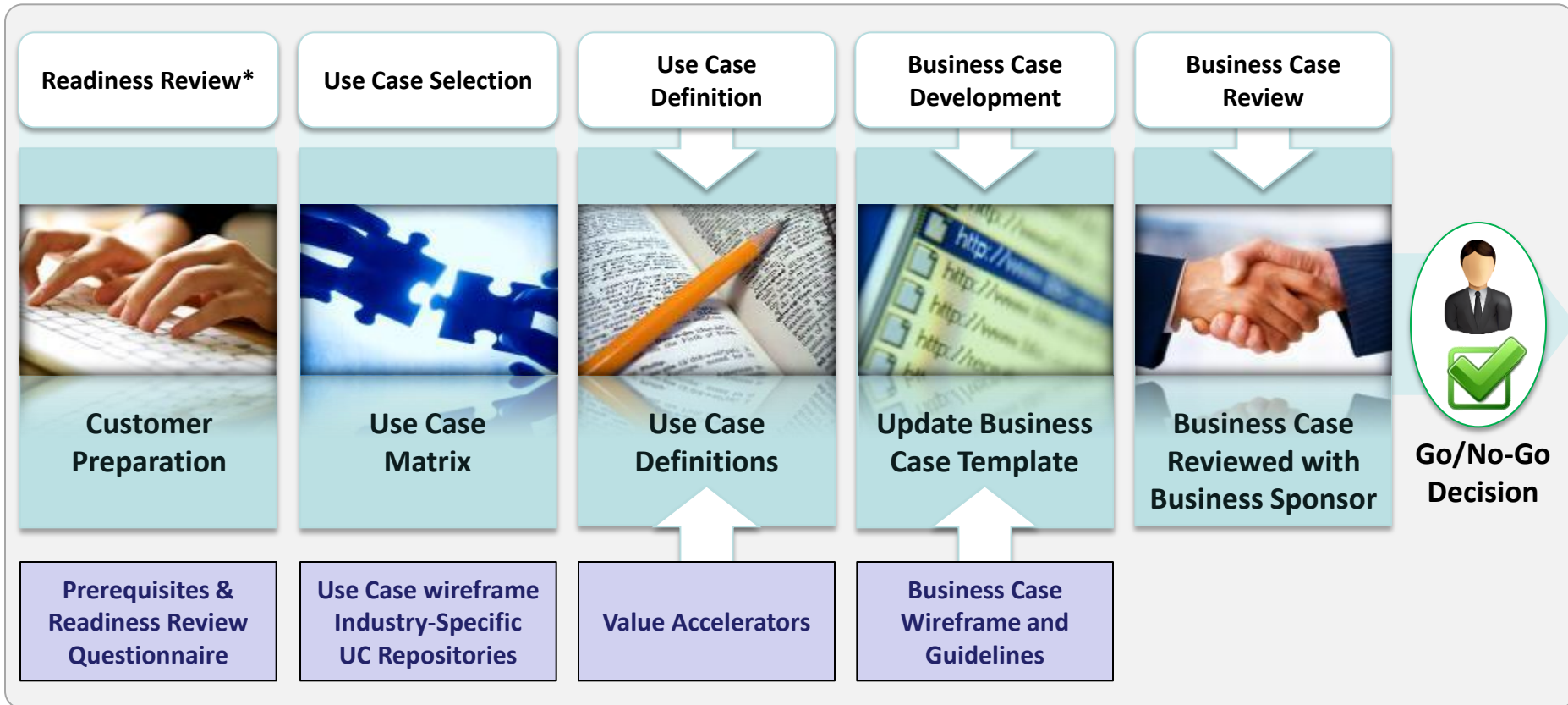
Build by layer

- Big Data takes in VVCV data
- Data processes quickly
- Analytics queries output meaningful insights
- Data is stored



Project: Big Data Business Case

4 weeks – Business Case Development

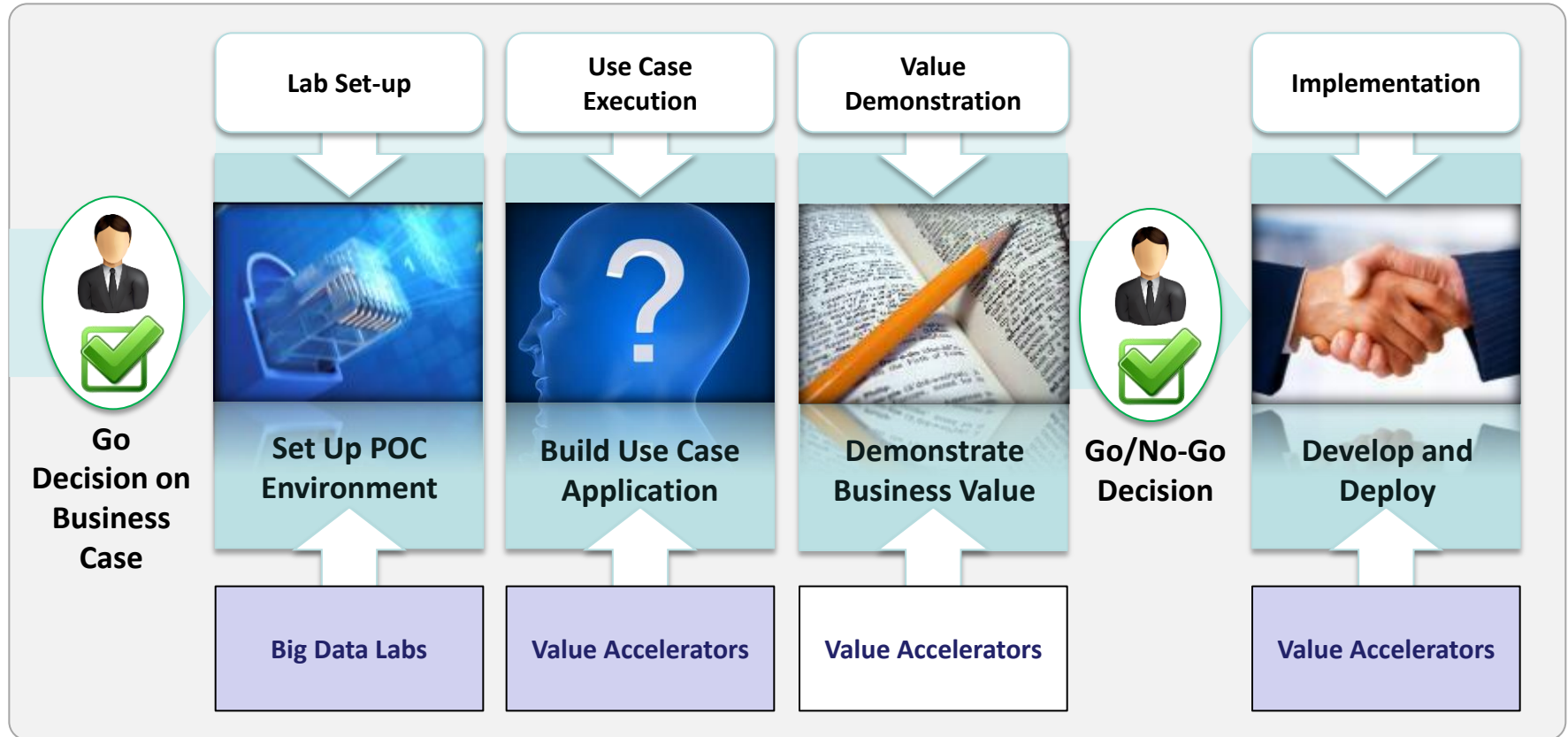


* Tasks outlined in red may be an investment and not client-billable.



Project: Big Data Proof of Concept

2+ weeks – Clear Scope of Use Cases and Complexity

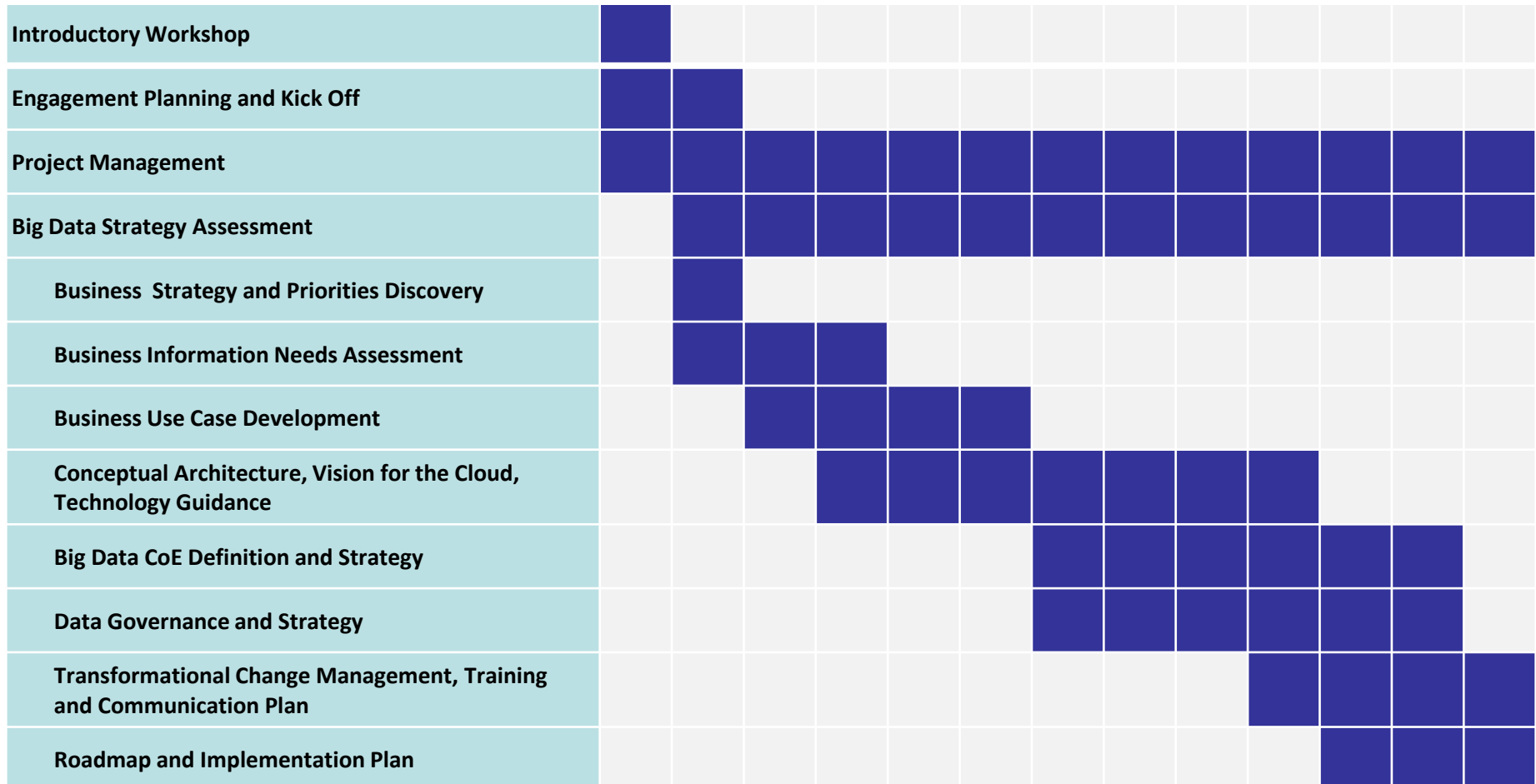




Project: Big Data Strategy Assessment

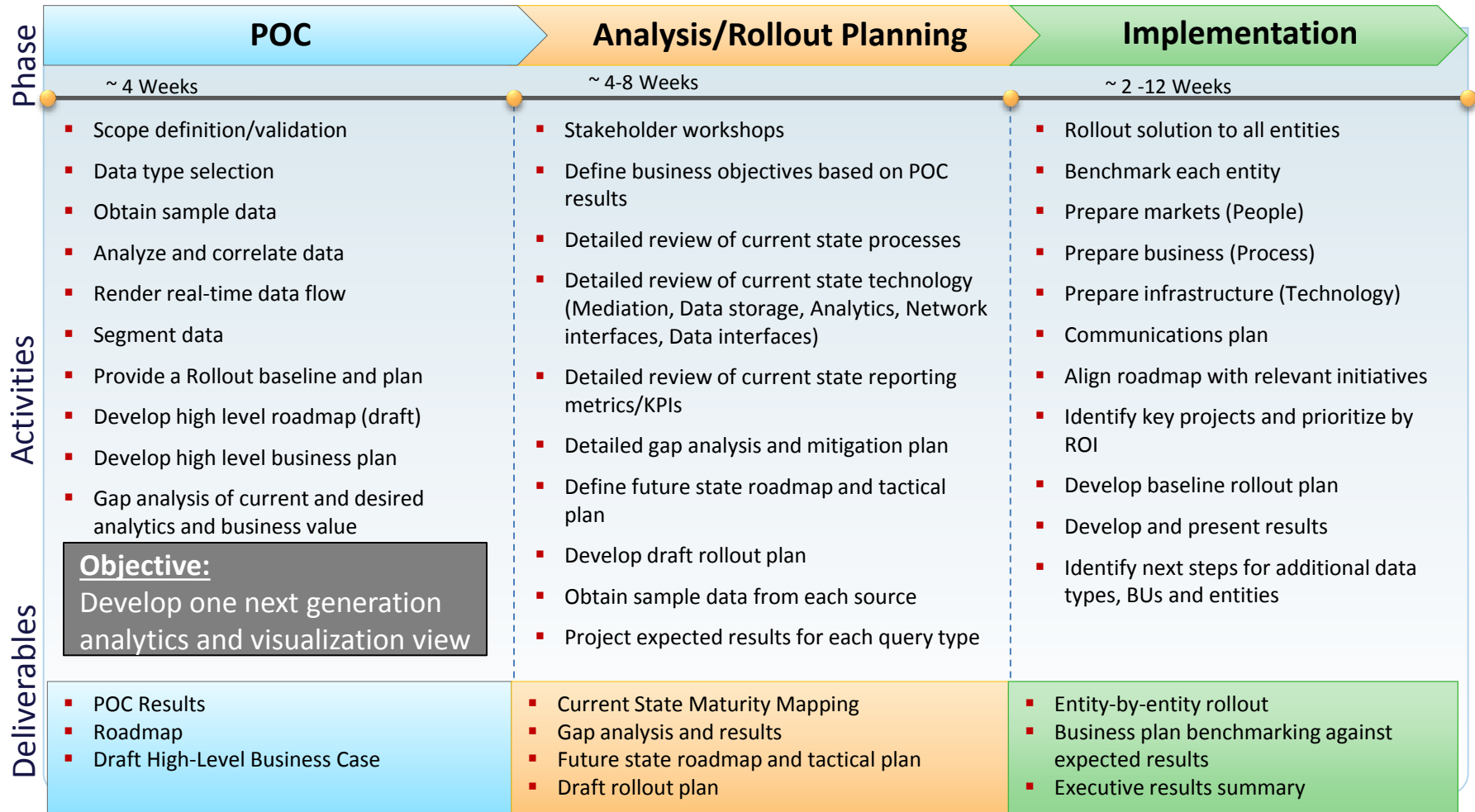
Project Activities/Week

0 1 2 3 4 5 6 7 8 9 10 11 12





Project: Big Data POC + Implementation





Unique Considerations

Data governance

- Data architecture
- Degree of centralization and decentralization
- New BU use case evaluation
- Metadata
- Deduplication
- Reconciliation
- Transitory data and backup
- Disaster recovery
- Site-to-site failover



Unique Considerations

Security and privacy

- Encryption
- Data masking
- Server locations
- Data retention
- Customer data requests



Unique Considerations

Services worth having

- Email validation
- IP validation
- GPS (latitude and longitude)
- IP location lookup
- Audit history
- Externalized metadata



Unique Considerations

Application process monitoring

Network monitoring and automation

Test automation

- Regression testing
- Required manual testing for credit cards, etc.
- Required manual testing for mobiles, locality, etc.

Compliance

Metrics and KPIs



Growth and Sustainability

How can systems adapt, streamline and grow?

- New data types and workflows
- Cluster management and node forecasting
- Automated deployment, network and application process monitoring
- Replay, backup, recovery and failover
- Cloud
- Social media and mobile integration
- SDN (software discovers needs and provisions the system)



Growth and Sustainability

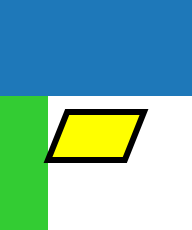
How can we frame our thinking to optimize business value?

- What are we trying to accomplish? How is this valuable?
- What data questions do we have now? What questions may follow?
- Can the opportunity or problem benefit from Big Data Analytics?
- Does the solution have growing room?



Hands-On Exercise

- Create a repeatable model for a client's Big Data Analytics in a focus area that inspires you and your team:
 - Vertical-specific system
 - Operations and support systems
 - Networking and infrastructure
 - Applications and integration
 - Research projects
 - Investments
 - Regulatory compliance
 - Adaptive language and distance communications
 - People and social life
 - Marketing and advertising
 - Education, employment and recruitment
- Continue this exercise and discussion via social media and online applications.



References

Search online using the terms and logos in this presentation.

Big Data Analytics Project Management book by Tiffani Crawford – <http://www.amazon.com> and search by title and author

Hadoop – <http://www.hadoopilluminated.com/>

NoSQL – <http://nosql-database.org/>

AWS Ecosystem – <http://aws.amazon.com/>

Data Science – <http://www.datasciencecentral.com/>

Agile – <http://agilemanifesto.org/>

Scrum – <http://www.scrumalliance.org/>



Speaker and Contact Information

Tiffani Crawford, PhD

408-829-7096

zapquick@aol.com

Big Data Analytics Project Management book by Tiffani Crawford – See Amazon Books at <http://www.amazon.com> and search by title and author.

Tiffani Crawford, PhD, builds global Big Data Analytics systems. She has 20 years of high technology experience with Fortune 500 companies, including Cisco Systems, Cognizant, Bank of America, VISA/Inovant, BAE Systems, Applied Competitive Technologies, Ditech Networks/Nuance, Big 4 financial firms, defense contractors and startups. She has worked in seminal technology development in Big Data, analytics, cloud, networking, telecommunications, software development, distributed multi-tier applications, multimedia/digital, geographic information systems, intelligent transport systems, finance, security, policy systems and structural equation modeling. She is a credited software developer and published author. She earned her PhD from the University of Southern California in 2005. She has also earned her Master's, Bachelor's and various technology certifications. She is a member of PMI with various philanthropic contributions.