THINKBIG ANALYTICS

Big Data Beyond the Hype: Trends, Pitfalls and Building Competency

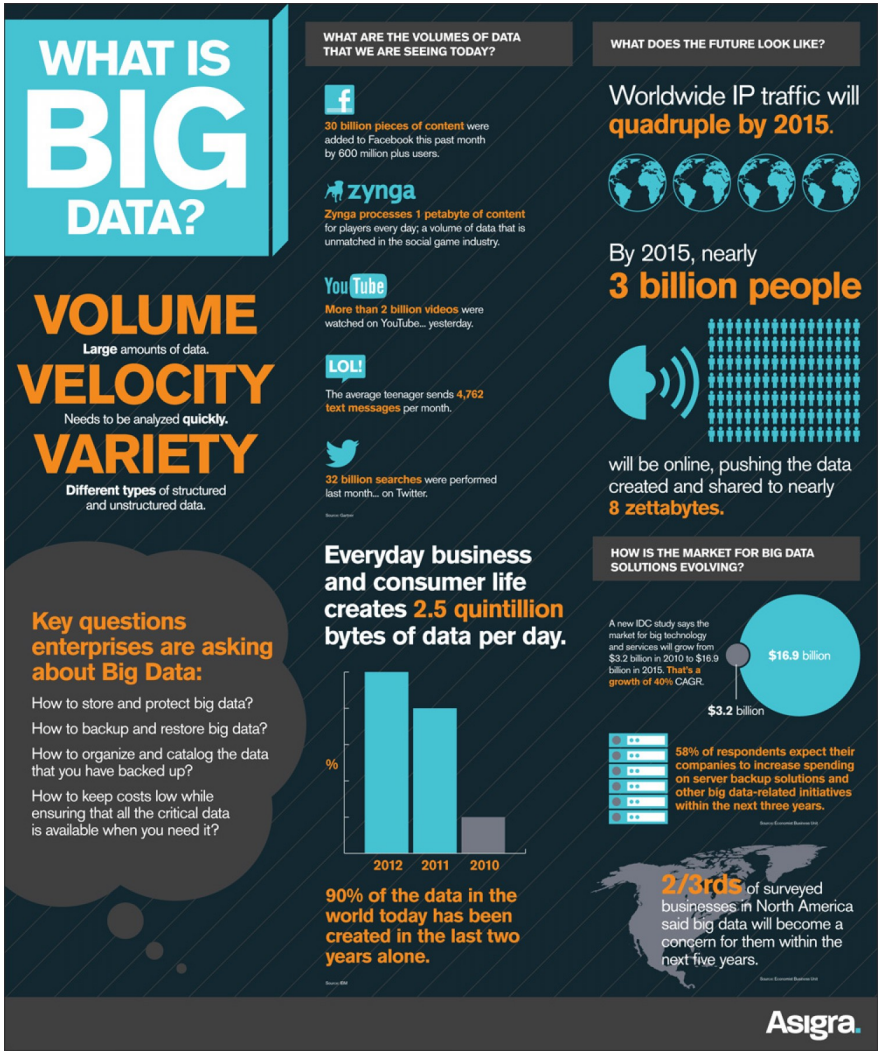
May 2014

Agenda

- Why Big Data
- Technology Capabilities
- Organizational Evolution
- Case Study
- Conclusions

THINKBIG ANALYTICS

Big Data



- *Data sets so large and complex that they become awkward to work with using standard tools and techniques*
- Google, Quantcast, Yahoo, LinkedIn… customer innovation in open source

Patterns from Delivering Big Data at Scale

eCommerce
2 of Global Top 5

Retail
2 of Global Top 5

Social Networking
Global #1

Credit Issuer
2 of Global Top 5

Banking
4 of Global Top 10

Financial Data Services
2 of Global Top 5

Financial Exchanges
Global #2

Internet Transaction Security
Global #1

Brokerage & Mutual Funds
2 of Global Top 5

Asset Management
Global #1

Semiconductor
2 of Global Top 5

Data Storage Devices
3 of Global Top 5

Disk Drive Manufacturing
Global #1

Telecommunications
2 of Global Top 5

Media & Advertising
2 of Global Top 4

Manufacturing Analytics

1	**Improve Yield** - reduced scrap, faster time to market
2	**Ease data access** – quick search access to all relevant data for all history in one place, not samples, summary or recent data
3	**Process Efficiency** – identify redundant tests, optimize throughput
4	**Component Analysis** – proactive discovery of component problems, compatibility detection
5	**Service Revenue Generation** - Leverage customer support to generate new revenues by increasing customer retention, support renewals, and upgrading customers to new products and licenses.

THINKBIG ANALYTICS

Product Analytics

1	**Issue Resolution** - Faster resolution of issues through deep analysis of integrated device log, product, and configuration data across the install base.
2	**Support Metrics** - Gain new insights into dimensions impacting service metrics by combining product use technical data with business data. E.g., number of tickets resolved, length of calls, resource allocation, customer satisfaction.
3	**Proactive Service** - Identify and address potential issues rather than waiting to react to them after they happen. I.e., intelligence for condition-based maintenance.
4	**Customer Self-Service** - Drive higher rates of customer self-service by providing customers access to tools and resources to answer their own questions and solve their own issues by analyzing past customer activity and integrating technical details from customer systems into self-service portals.
5	**Service Revenue Generation** - Leverage customer support to generate new revenues by increasing customer retention, support renewals, and upgrading customers to new products and licenses.

THINKBIG ANALYTICS

Clickstream Analytics

1	**Timely Processing** - Data volumes are increasing clogging data warehouse, slow and out of date.
2	**Cost Efficiency & Scale** - Data volumes are increasing clogging data warehouse, slow and out of date.
3	**Flexible Tagging** – Analyze site without having to push data to tags, classify activity based on sections, URLs more flexibly than tags.
4	**Converged Analytics** - Understand and improve web, mobile, cross-channel customer activity. Siloed apps costly & inflexible – e.g., tags require feeding data about user demographics, segments. Mash up interactions with data like geographic, demographic, social, call center, purchases.
5	**Deeper Exploration** - Analyze sequences of events in customer journeys. Deep drill down (how do users with the latest iPhone and iOS compare to latest Samsung and Android). Cohort analysis - those who buy books vs sporting goods.
6	**Predictive Models** – next best offer recommendations, personalization, customer value, microsegmentation, ad attribution…

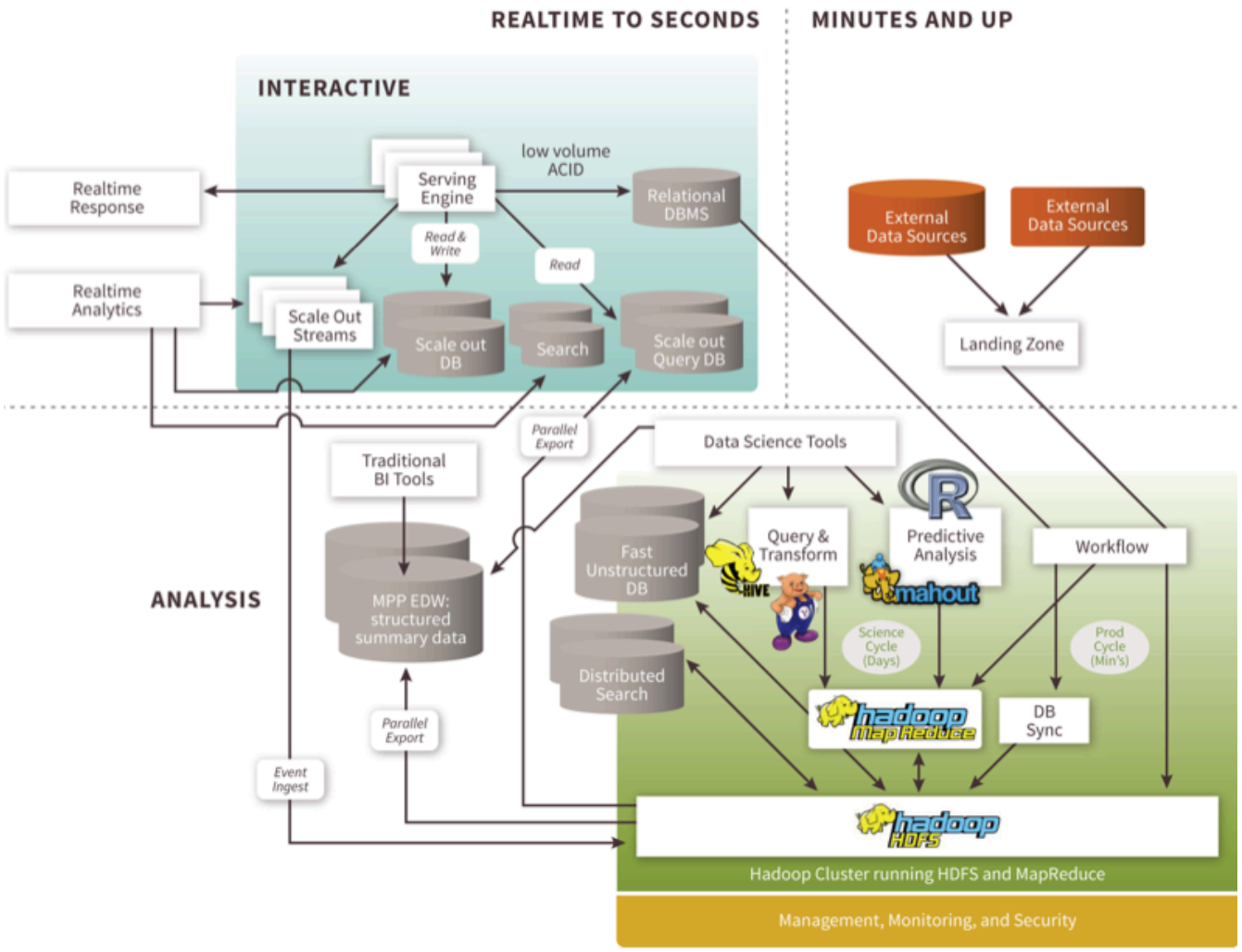
THINKBIG ANALYTICS

Agenda

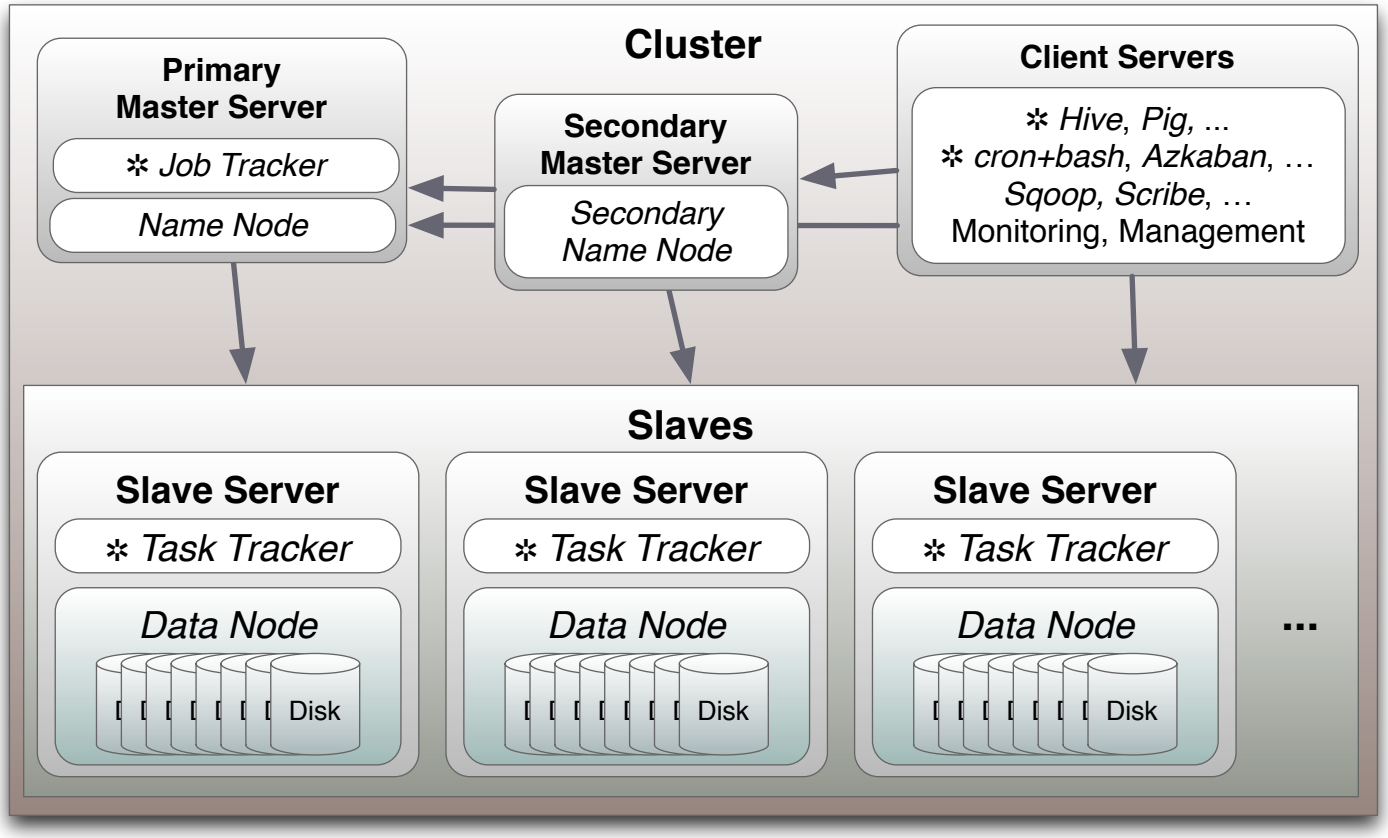
- Why Big Data
- **Technology Capabilities**
- Organizational Evolution
- Case Study
- Conclusions

THINKBIG ANALYTICS

Big Data Reference Architecture

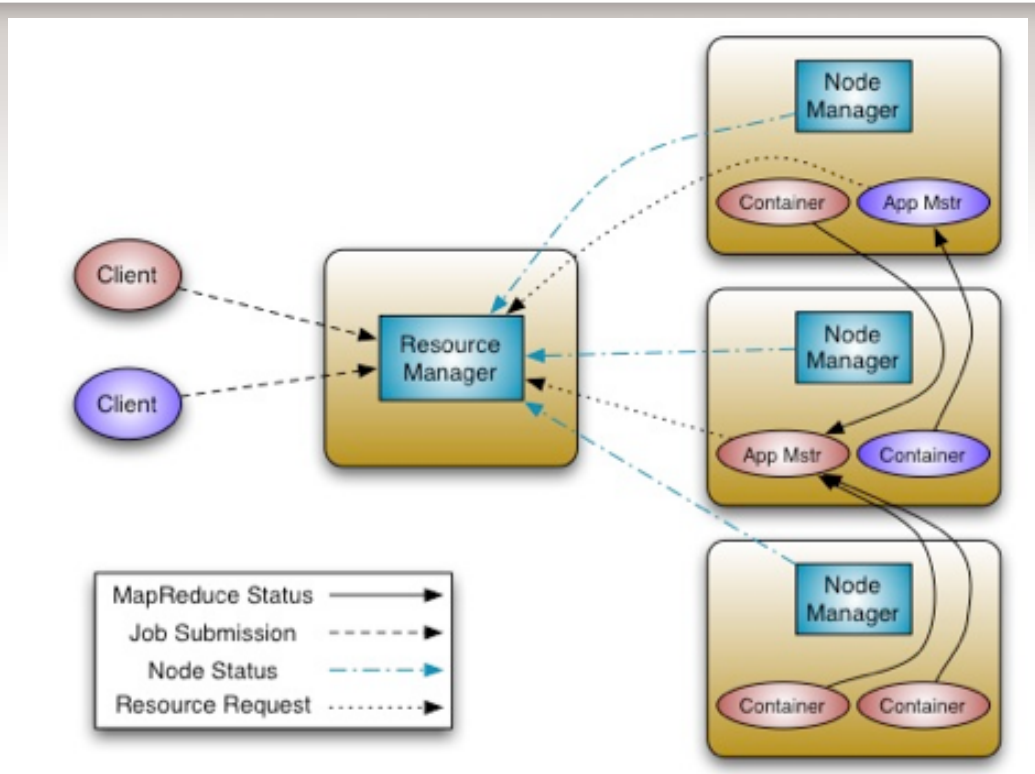


Apache Hadoop



- Community Open Source
- Over $1 Billion invested in startups, all major sw companies embrace
- 3 replicas for HA, 50 PB+ scale

YARN - Yet Another Resource Negotiator



- Support for many processing engines including emerging Predictive Analytics libraries
- Common cluster with local data access in HDFS

Narrowing In

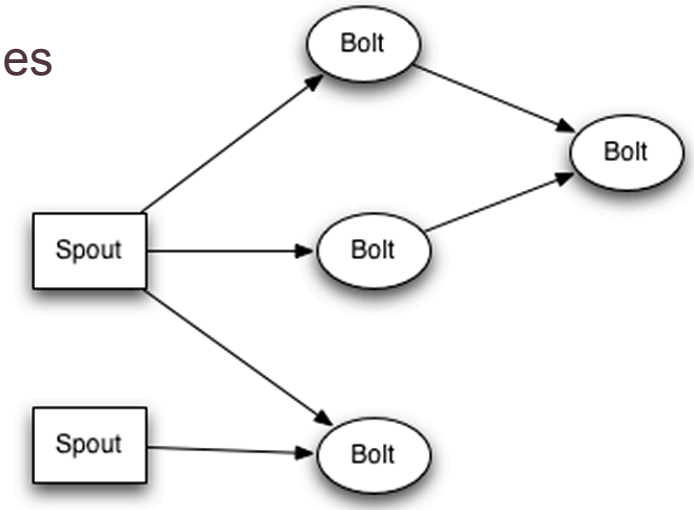
- Focus on technologies for prescriptive analytics
  - Getting insights from your data
- To get there you need
  - data ingestion (Flume, Kafka, Sqoop, ...)
  - data transformation (Pig, MapReduce, Hive, Cascading, …)
- Beyond prescriptive, Machine Learned predictive analytics is interesting

Apache Spark

- Scala-based more general distributed processing with long-lived memory caching
- First released 2010, many contributors
- Scala, Java, Python APIs
- Largest production cluster to date is 80 nodes at Yahoo
- YARN Support
- Subprojects
  - Shark – Hive on Spark
  - MLib – machine learning
  - Spark Streaming
  - GraphX
- Pros: faster iterative analysis, growing distribution support
- Cons: less mature, Scala API foundation, smaller community

Apache Storm

- DAG Processing of never ending streams of data
  - First released 2011
  - Used at Twitter plus dozens of other companies
  - Reliable - At Least Once semantics
  - Think MapReduce for data streams
  - Java / Clojure based
  - Bolts in Java and ‘Shell Bolts’
  - Not a queue, but usually reads from a queue.
  - YARN integration



- Pros: community, HW support, low latency, high throughput
- Cons: static topologies & cluster sizing, Nimbus – SPOF, state management

Platform Trends

- Hadoop will eat Big Data Analytics
- Hadoop YARN to run diverse applications
- Realtime latency: Storm, Samza, HBase, Cassandra, Elasticsearch…
- Management: Ambari, Pepperdata…
- Security: Knox, Sentry, XA Secure…
- Cloud: AWS, Qubole, Altascale, Info Chimps…

Analytics Trends

- Realtime query engines: Hive/Tez, Shark, Impala, Presto…
- Machine Learning: Spark, MLib, Mahout, Orryx…
- Metadata: HCatalog
- Analyst Tools: Tableau, Alpine Data Labs, Platfora, Zoomdata…
- Sensemaking: Data Tamer, Trifacta, Paxta…

Agenda

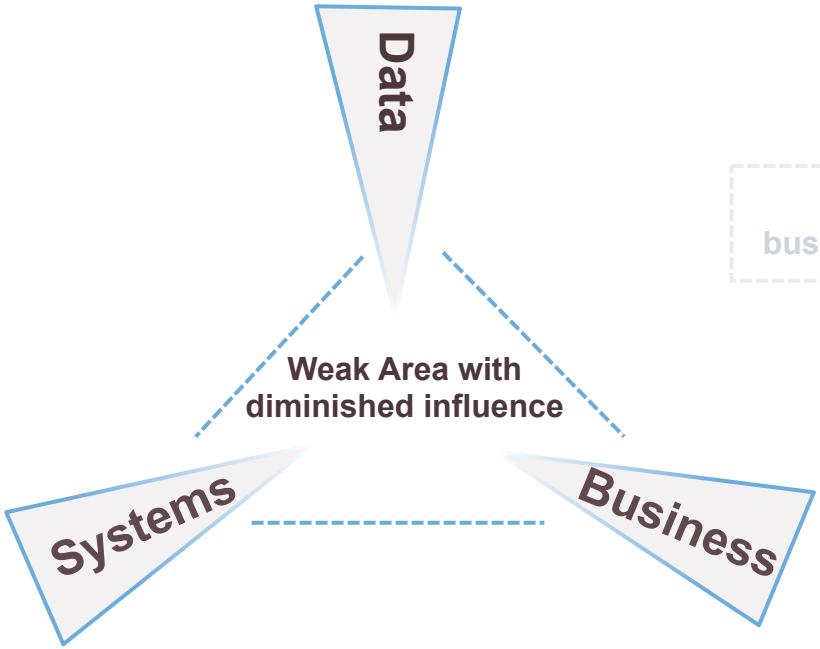
- Why Big Data
- Technology Capabilities
- **Organizational Evolution**
- Case Study
- Conclusions

THINKBIG ANALYTICS

How To Get the Right Answers

Current State

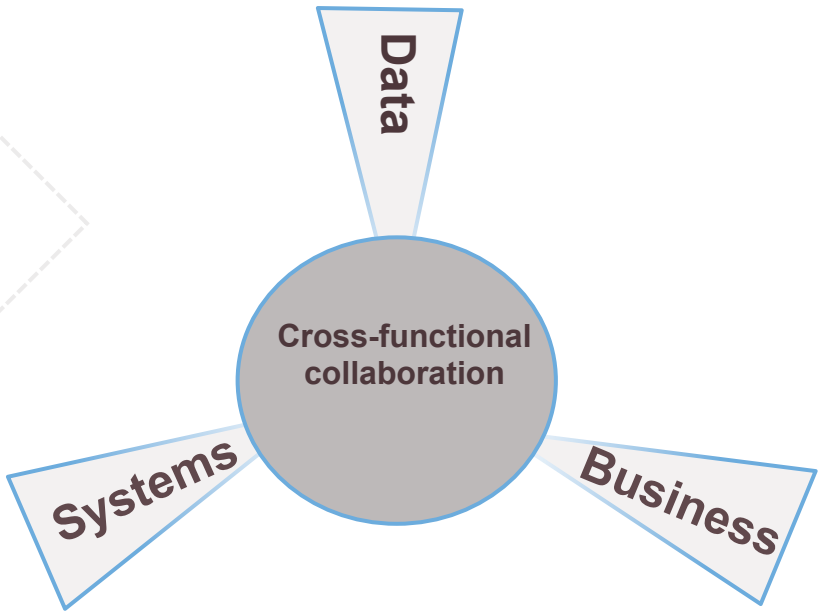
Weak intersection between Data, Systems and Business often means project failure



Data-driven business alignment

Improved State

Collaborative integration of Analytics with Big Data capabilities for success.



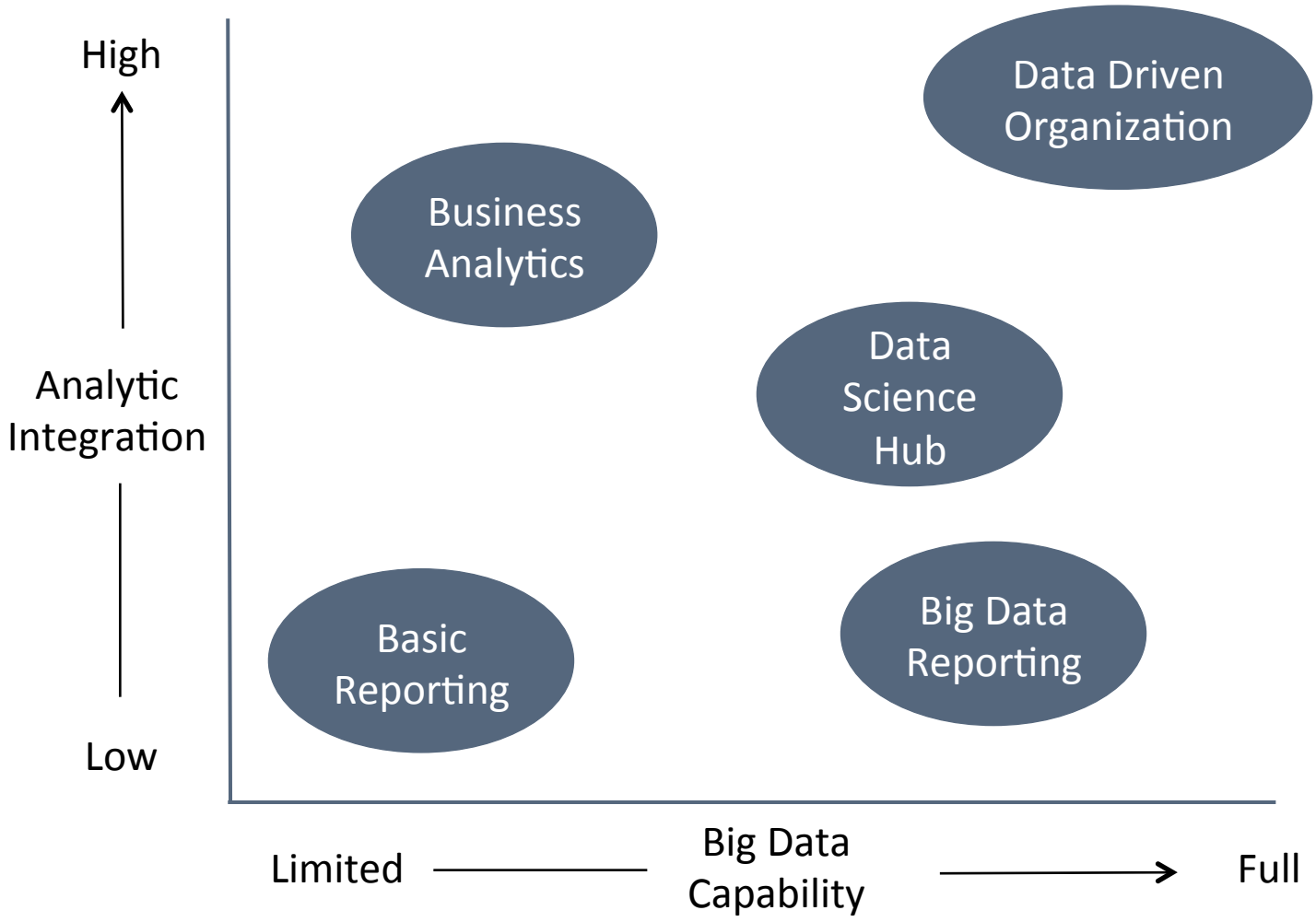
Analytics and Data Science

Analytics is the umbrella term for discovery and communication of signal in data.

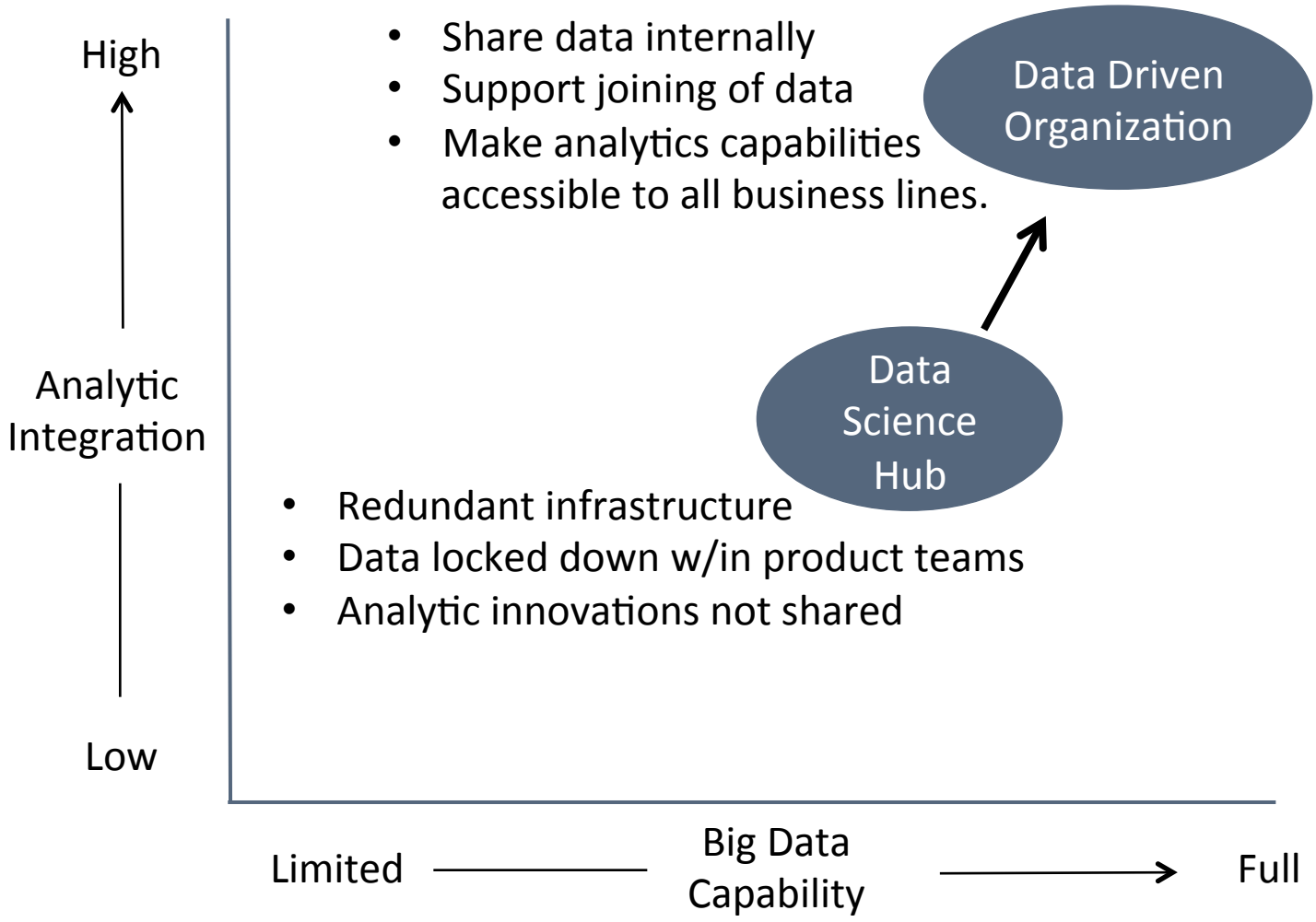
- *Intelligence and Reporting*: Governed insights for general audiences by providing summaries and visualizations over well understood data
- *Descriptive Analytics*: Interpretable data stories to domain consumers (e.g. business, sales, etc.) using exploratory data analysis and descriptive models (e.g. clustering) over well understood data
- *Predictive Analytics*: Predicts future state or assesses impact of changing parameters for domain consumers (e.g. business, sales, etc.) using statistical modeling over well understood data
- *Data Science*: Uncovers new or missing signals and relationships in data to be leveraged by other analytics. Supported with a mix of exploratory data analysis and statistical modeling over unexplored data or new combinations of known data.
  New capability driven by Big Data, currently overhyped and overused.

End-goal: effective and repeatable integration of analytics into business processes.

Analytics Capability Model



Client Example – Internet Infrastructure Firm



Agenda

- Why Big Data
- Technology Capabilities
- Organizational Evolution
- **Case Study**
- Conclusions

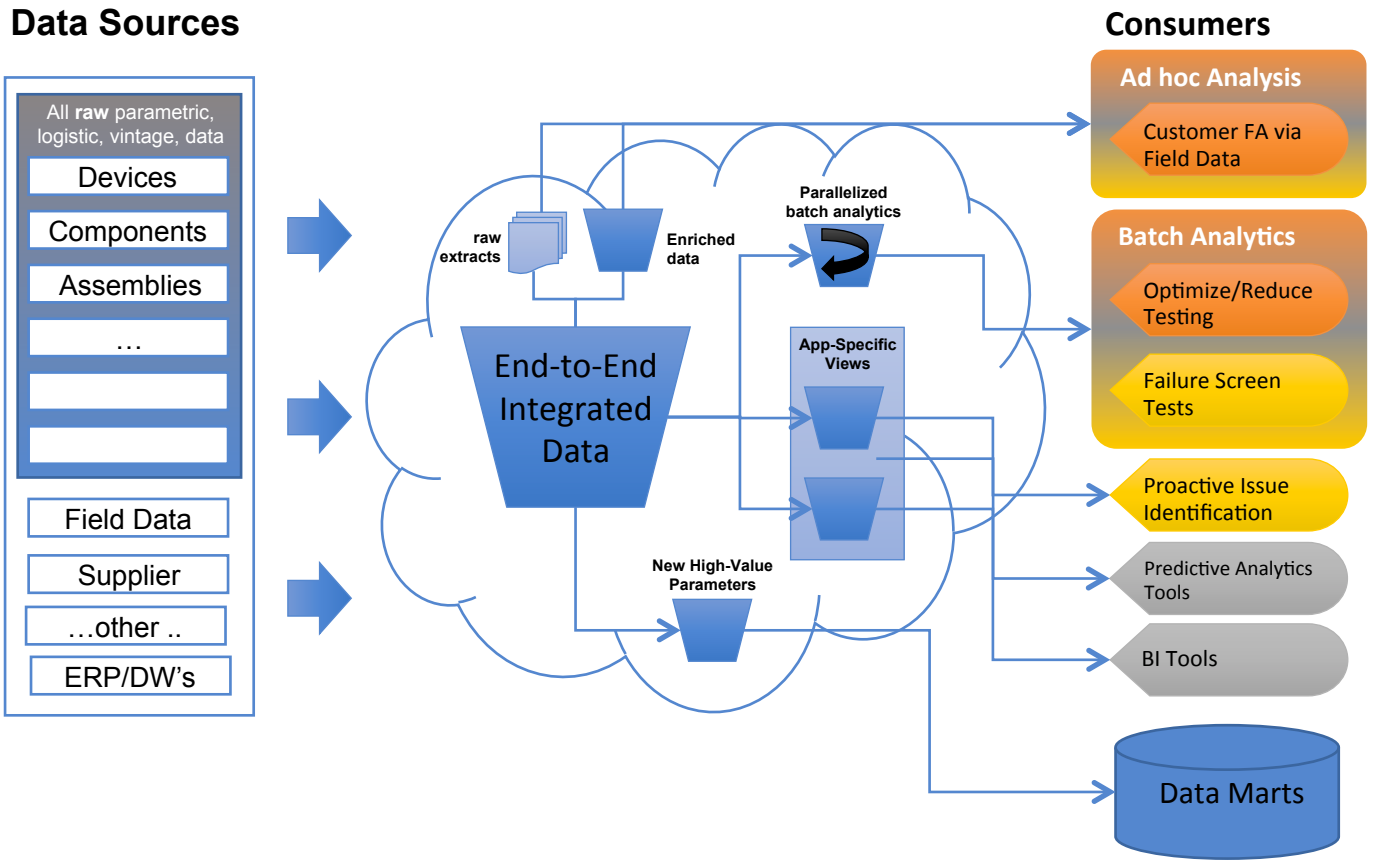
THINKBIG ANALYTICS

Use Cases

- "Reduce Data Search Parties"
  - Stop playing "Where's Waldo with your Data"
    - " I know I have that data….. somewhere?"
  - Data Aggregation to a Common Platform with common access tools

- Improve Yields by Accessing More Data in a More Timely Manner
  - End to end visibility for every test, every diagnostic and all info from all components of a product (internal and external)
  - Speed up yield improvement ramp up on new products
  - Improve steady state yield on existing products.

Big Data Platform



Highlights

- Enterprise-wide data access for timely analytics and insights
- Foundation for large scale proactive analytics

	Legacy	BDP
Retention	3-6 months scattered DBs then tape archive	All data online in common data lake for 3+ years
Coverage	Summaries, samples for several data sets	All parametric data captured in raw and integrated form
Analysis	Reactive resulting in missed improvement opportunities	In daily operations and on larger data sets to support proactive improvements

THINKBIG ANALYTICS

Key Metrics & Highlights

- Metrics:
  - Collecting **>2M** manufacturing/testing binary files daily
  - Collecting from ~**500** tables across 6 databases → **tens of millions** of records daily
  - Over **140** users to date in early piloting
  - Over **150** attendees participated in BDP training
- Highlights: although early in the overall journey, the BDP is already demonstrating early benefits:

DATA SEARCH PARTIES

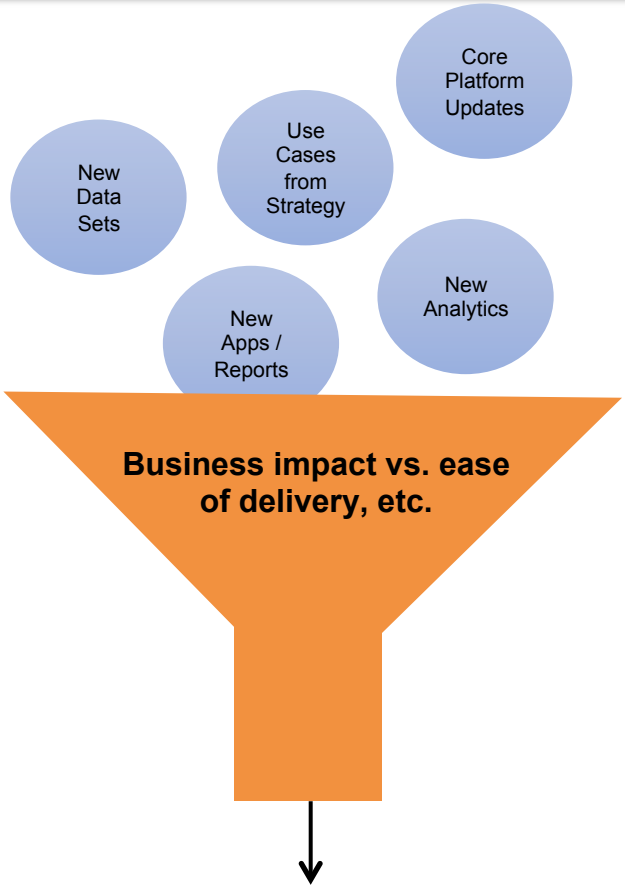
Development Engineer: obtained technical data **from the BDP in hours as opposed to 3+ weeks** to pull from tape archive

Development Engineer*:* demonstrated the joining of data sets for detailed logistics tracking—**analyses that is very difficult to conduct with current systems**

YIELD

Ops Engineer**:** a recent production issue required detailed historical data. Current systems did not have the required retention for this data. However, the team was able to pull the data from the BDP **in minutes, as opposed to 3+ weeks** to pull the data from tape archive

Governance – Project Prioritization

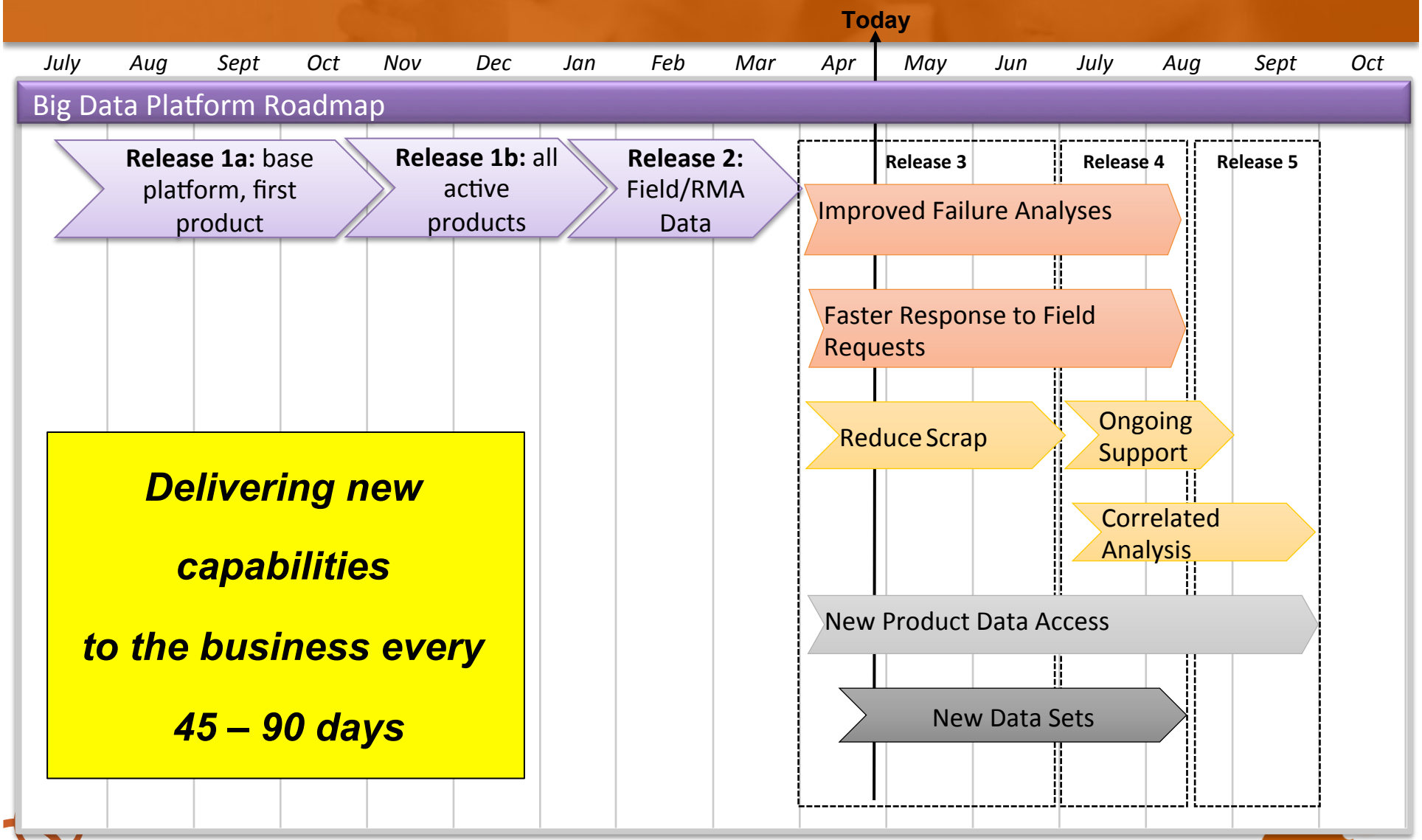


Nominated by… — **Business Experts**

Prioritized by… — **Working Steering Committee**

Sponsored by… — **Executive Steering Committee**

Release Cycles

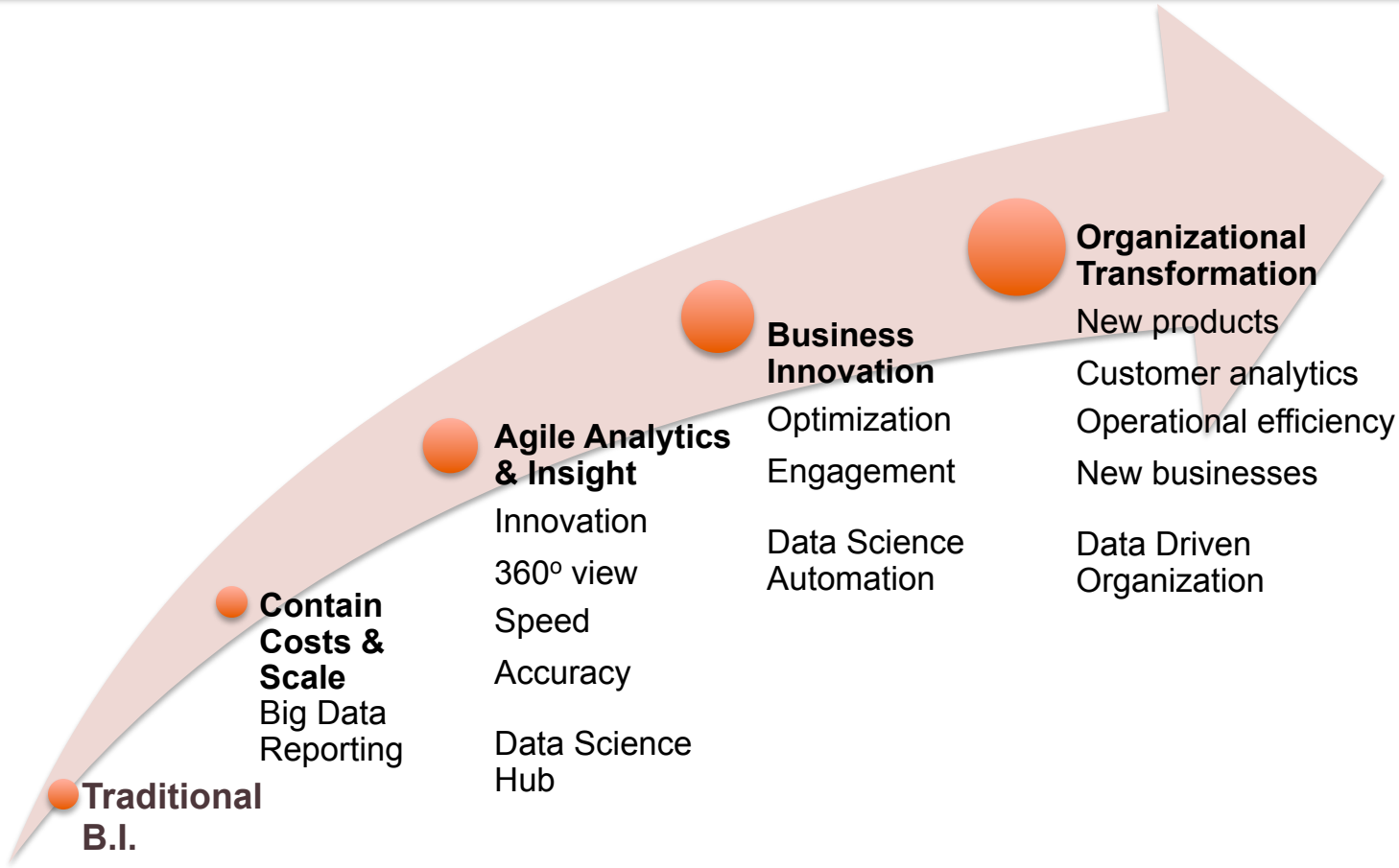


Agenda

- Why Big Data
- **Technology Capabilities**
- Organizational Evolution
- Case Study
- **Conclusions**

THINKBIG ANALYTICS

Stages of Big Data Adoption



Sustained competitive advantage through Big Analytics

Pitfalls

- Data Politics
- Immature Governance
- Siloed Organization
- Siloed Applications
- Buy-in
- Leaping over phases / Bypassing low hanging fruit
- Disruptive Change
- Insufficient funding
- Big Bang Rollout
- Skills Gap
- Tactical use cases/the big data plateau
- Business as usual

Conclusions

- Big Data is affecting all companies especially in tech
- Start with low hanging fruit
- Establish a roadmap for incremental organizational evolution and technical capabilities

THINKBIG ANALYTICS