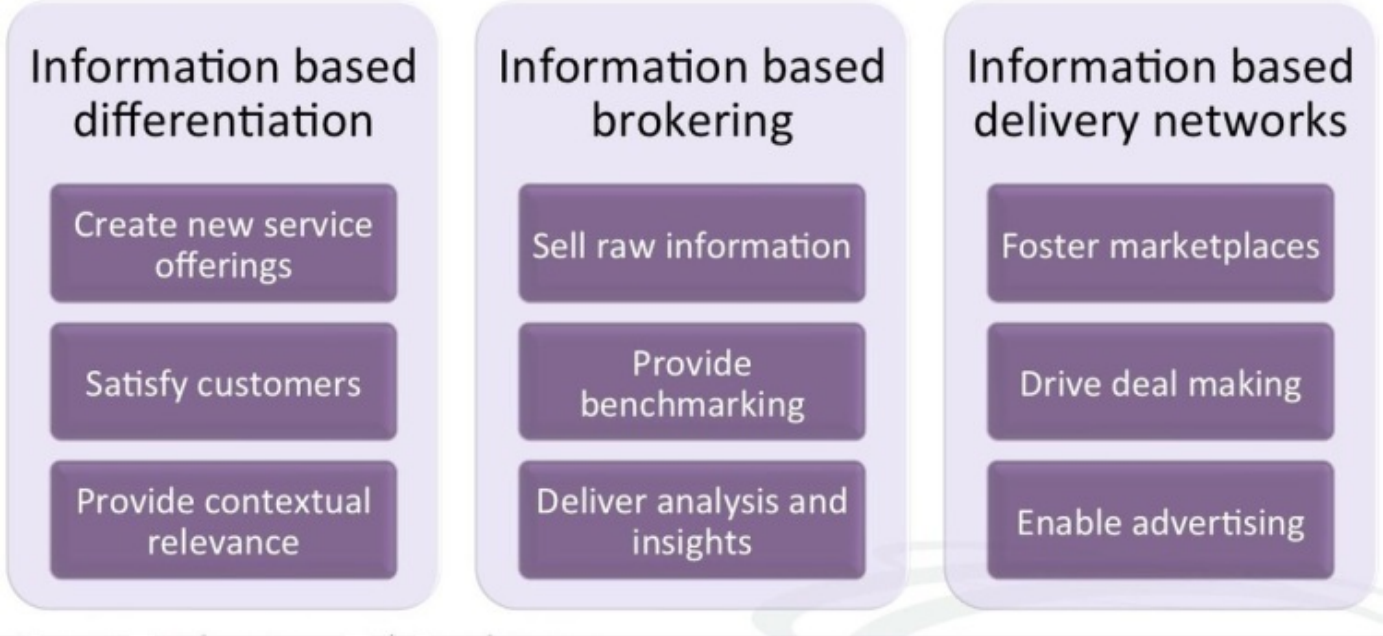
Michael Walker's Blog (80)

Data Science Ethics & Big Data Business Models

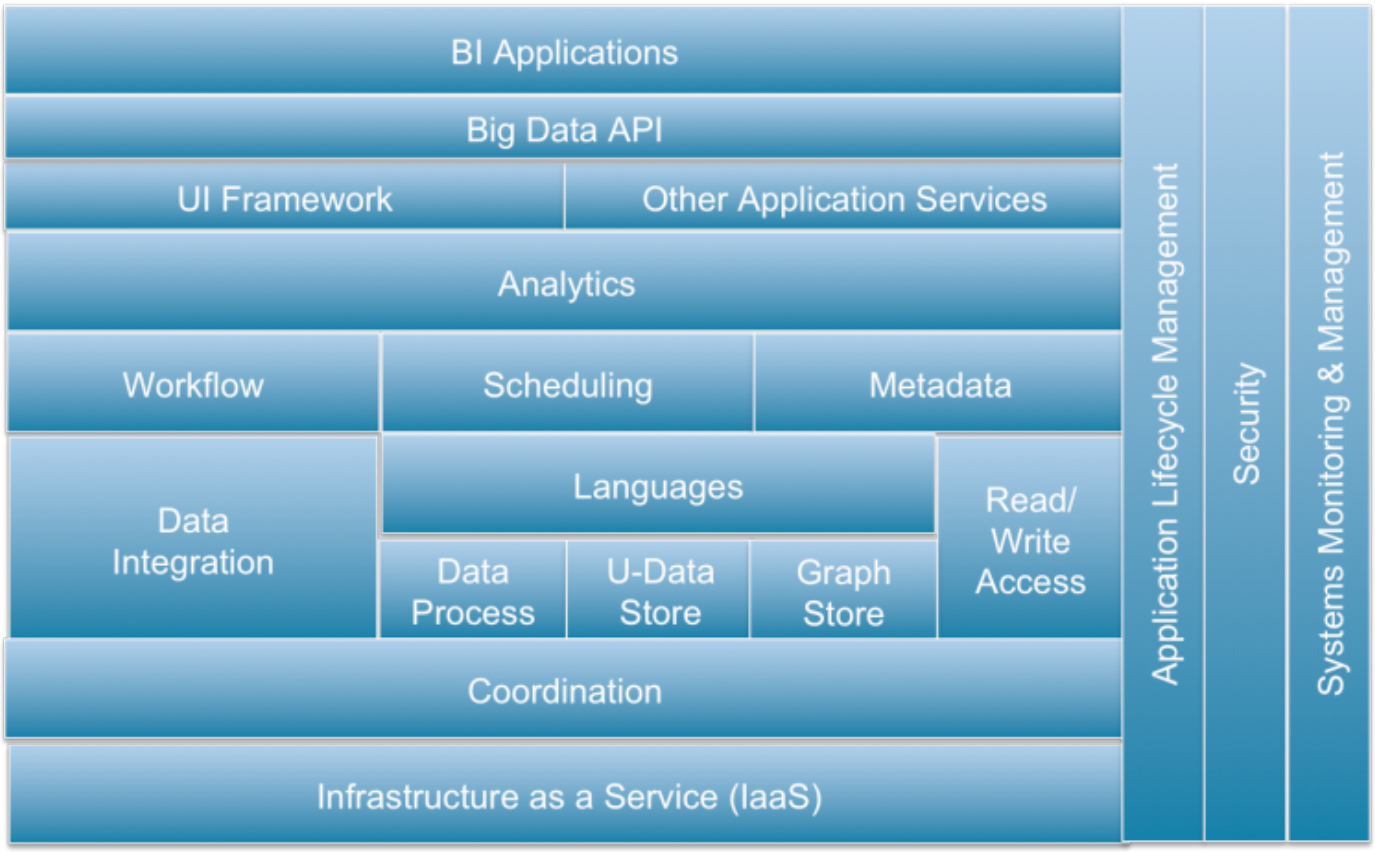


Ray Wang's HBR piece "...

Continue

Added by Michael Walker on January 10, 2013 at 9:30am — No Comments

Big Data Platforms as a Service

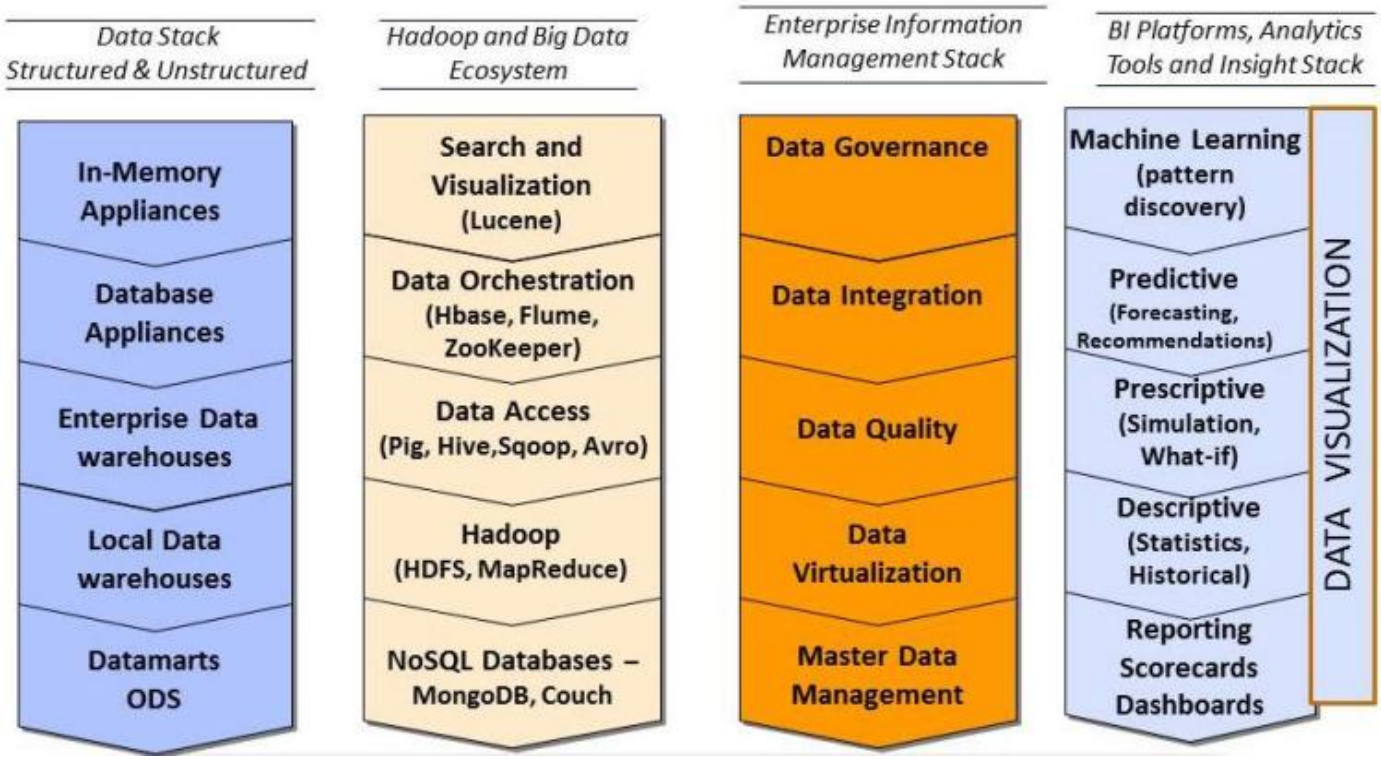


Big Data Platforms as a Service (PaaS) lets an organization take advantage of a service providers compute power, analytical tools, store as...

Continue

Added by Michael Walker on January 2, 2013 at 8:46am — 1 Comment

Big Data Analytics Infrastructure



Recent surveys suggest the number one investment area for both private and public organizations is the design and building of a modern data...

Continue

Added by Michael Walker on December 26, 2012 at 8:11am — 2 Comments

Structured vs. Unstructured Data: The Rise of Data Anarchy

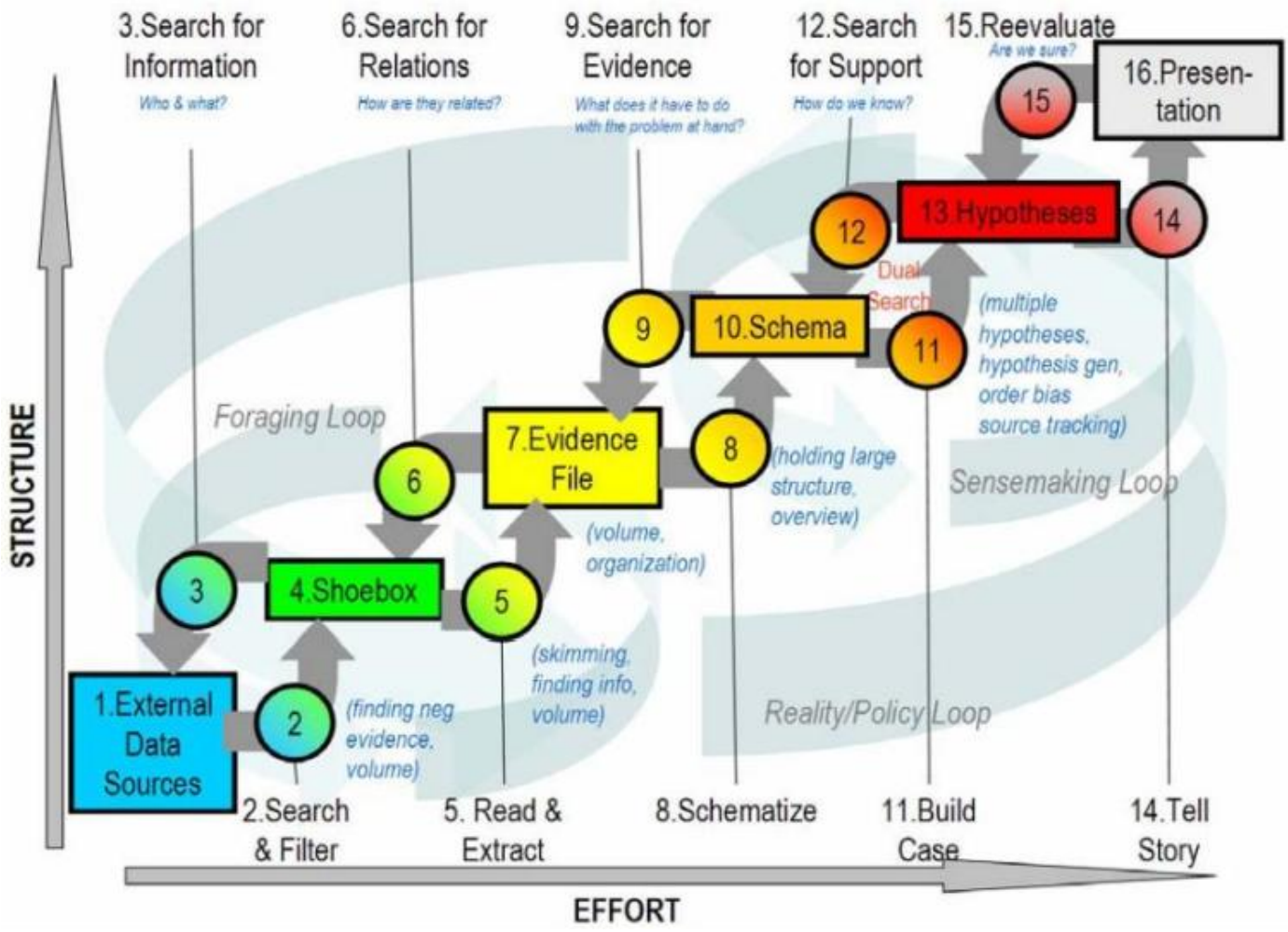
Data science and business analytics works with both structured and unstructured data. Yet the future belongs to unstructured or semi-structured data from both internal and external sources.

Total Enterprise Data Growth 2005-2015...

Continue

Added by Michael Walker on December 19, 2012 at 9:30am — No Comments

Data Science for Better Business Decisions Formula



After designing and building a modern data warehouse / business intelligence / data analytical ecosystem, many clients are frustrated they are...
Continue

Added by Michael Walker on December 12, 2012 at 2:30pm — No Comments

Three Stages of Analytics Adoption

THE THREE STAGES OF ANALYTICS ADOPTION

Three capability levels — Aspirational, Experienced and Transformed — were based on how respondents rated their organization's analytic prowess.

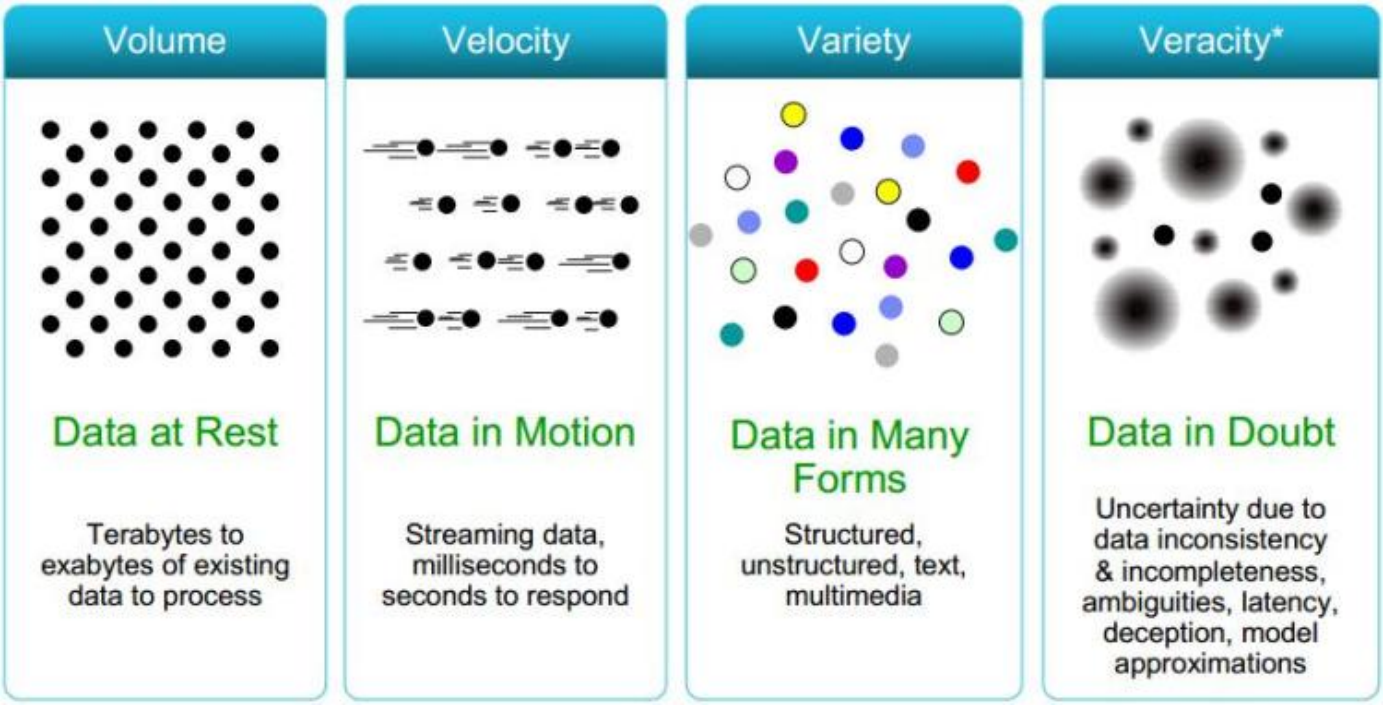
	ASPIRATIONAL	EXPERIENCED	TRANSFORMED
Motive	•Use analytics to justify actions	•Use analytics to guide actions	•Use analytics to prescribe actions
Functional proficiency	•Financial management and budgeting •Operations and production •Sales and marketing	•All Aspirational functions •Strategy/business development •Customer service •Product research/development	•All Aspirational and Experienced functions •Risk management •Customer experience •Work force planning/allocation •General management •Brand and market management
Business challenges	•Competitive differentiation through innovation •Cost efficiency (primary) •Revenue growth (secondary)	•Competitive differentiation through innovation •Revenue growth (primary) •Cost efficiency (secondary)	•Competitive differentiation through innovation •Revenue growth (primary) •Profitability acquiring/retaining customers (targeted focus)
Key obstacles	•Lack of understanding how to leverage analytics for business value •Executive sponsorship •Culture does not encourage sharing information	•Lack of understanding how to leverage analytics for business value •Skills within line of business •Ownership of data is unclear or governance is ineffective	•Lack of understanding how to leverage analytics for business value •Management bandwidth due to competing priorities •Accessibility of the data
Data management	•Limited ability to capture, aggregate, analyze or share information and insights	•Moderate ability to capture, aggregate and analyze data •Limited ability to share information and insights	•Strong ability to capture, aggregate and analyze data •Effective at sharing information and insights
Analytics in action	•Rarely use rigorous approaches to make decisions •Limited use of insights to guide future strategies or day-to-day operations	•Some use of rigorous approaches to make decisions •Growing use of insights to guide future strategies, but still limited use of insights to guide day-to-day operations	•Most use rigorous approaches to make decisions •Almost all use insights to guide future strategies, and most use insights to guide day-to-day operations

An article from MIT entitled "...

Continue

Added by Michael Walker on December 5, 2012 at 9:00am — No Comments

Data Veracity



Data Veracity, uncertain or imprecise data, is often overlooked yet may be as important as the 3 V's of...

Continue

Added by Michael Walker on November 28, 2012 at 3:00pm — No Comments

Data as Strategic Asset

A companywide approach

To help create a companywide approach to analytics, it is helpful to see which groups—if any—need common data to answer these six key analytical questions. If you find they do need common data, then it will make sense to set up systems and processes to ensure the groups are able to share the data.

	Past	Present	Future
Information	What happened? (Reporting)	What is happening now? (Alerts)	What will happen? (Extrapolation)
Insight	How and why did it happen? (Modeling, experimental design)	What's the next best action? (Recommendation)	What's the best/worst that can happen? (Prediction, optimization, simulation)

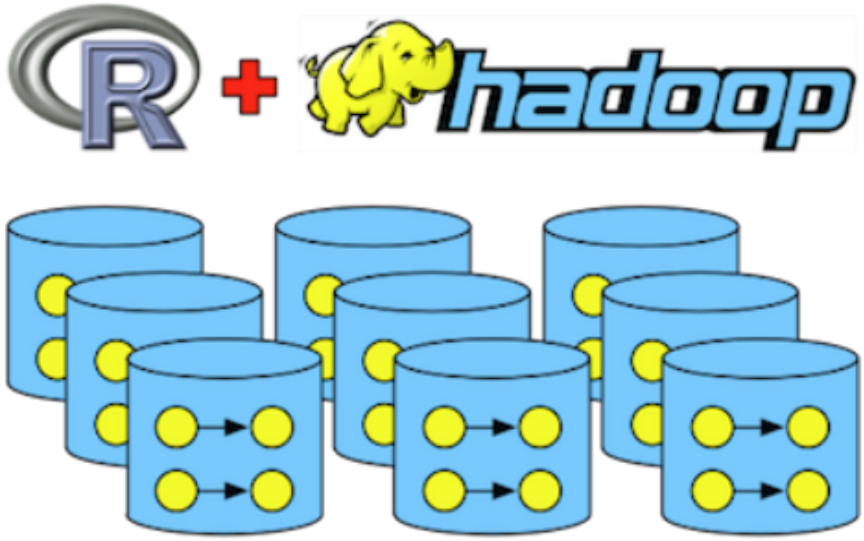
Transforming into a data-driven organization - turning information into actionable insights is a three (3) part strategy:

• Technology –...

Continue

Added by Michael Walker on November 14, 2012 at 9:22am — No Comments

R + Hadoop = Data Analytics Heaven

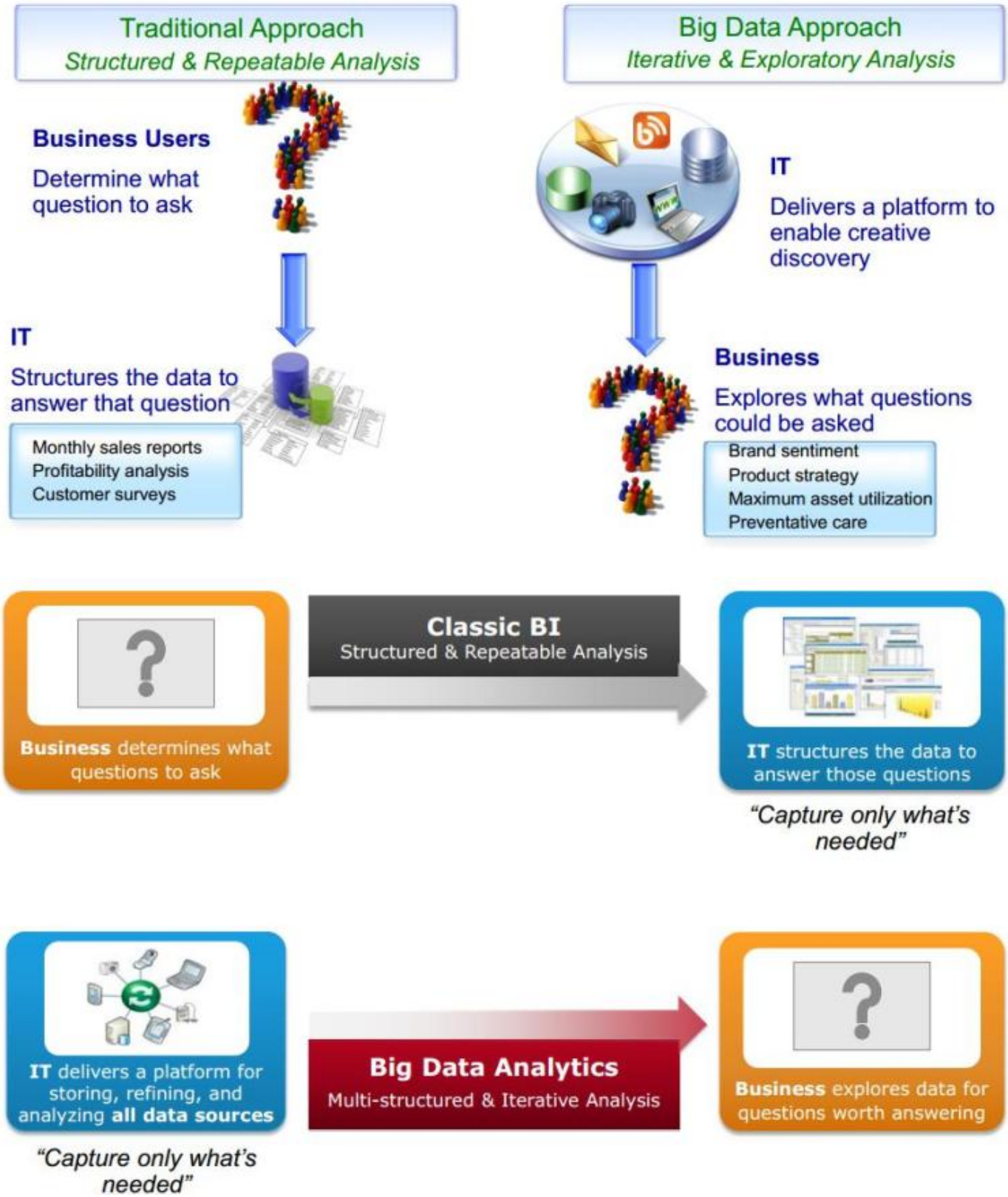


Hadoop (MapReduce where code is turned into map and reduce jobs, and Hadoop runs the jobs) is the most well known technology used for "Big Data" because it allows...

Continue

Added by Michael Walker on November 7, 2012 at 3:57pm — No Comments

Traditional BI vs Data Analytics Approach



Continue

Added by Michael Walker on October 31, 2012 at 7:38am — 1 Comment

Row vs Columnar vs NoSQL Databases

	Row-Based	Columnar	NoSQL—Key Value Store	NoSQL—Document Store	NoSQL—Column Store
Basic Description	Data structured in rows	Data is vertically striped and stored in columns	Data stored usually in memory with some persistent backup	Persistent storage for unstructured or semi-structured data along with some SQL-like querying functionality	Very large data storage, MapReduc support
Common Use Cases	Transaction processing, interactive transactional applications	Historical data analysis, data warehousing, business intelligence	Used as a cache for storing frequently requested data for a web app	Web apps or any app which needs better performance and scalability without having to define columns in an RDBMS	Real-time data logging as in finance or web analytics
Strengths	Capturing and inputting new records. Robust, proven technology.	Fast query support, especially for ad hoc queries on large datasets, compression	Scalability, very fast storage and retrieval of unstructured and partly structured data	Persistent store with scalability features such as sharding built in with and better query support than key-value stores	Very high throughput for Big Data, strong partitioning support, rando read-write access
Weaknesses	Scale issues—less suitable for queries, especially against large databases	Not suited for transactions; import and export speed; heavy computing resource utilization	Usually all data must fit into memory, no complex query capabilities	Lack of sophisticated query capabilities	Low-level API, inability to perform complex queries, high latency of response to queries
Typical Database Size Range		Several GBs to 50 TB	Several GBs to several TBs	Few TBs to several PBs	Few TBs to several PBs
Key Players	MySQL, Oracle, SQL Sever, Sybase ASE	Infobright, Aster Data, Sybase IQ, Vertica, ParAccel	MemCached, Amazon S3, Redis, Voldemort	MongoDb, Couchdb, SimpleDb	HBase, Big Table, Cassandra

See: http://bit.ly/RWBoCk...

Continue

Added by Michael Walker on October 24, 2012 at 2:43pm — 4 Comments

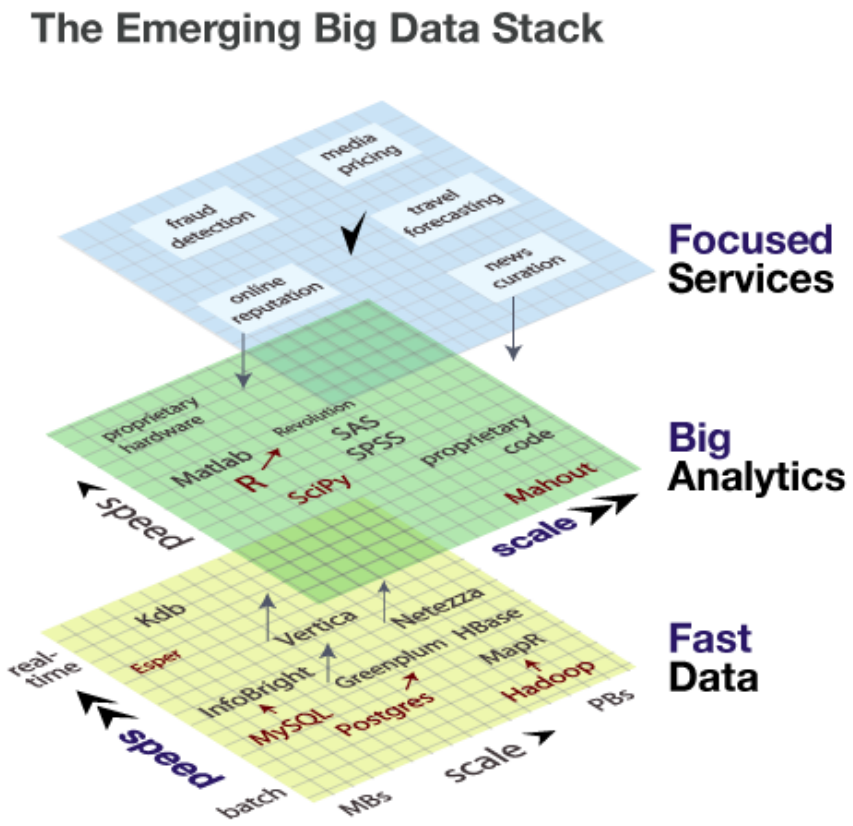
Gale-Shapley Algorithm

Al Roth along with Lloyd Shapley won the Nobel Prize for matching and the design of new types of markets. The Gale-Shapley algorithm is a cornerstone of the matching methods Al Roth pioneered. The algorithm has been extended by Roth and computer scientists including Don Knuth to apply "Matching Theory" to design...

Continue

Added by Michael Walker on October 17, 2012 at 1:30pm — No Comments

The Emerging Data Stack and Mobile Access



The emerging "Data Stack" or "Data Layer" is in full transition and can be viewed and defined many different ways. The ability to capture, analyze and learn...

Continue

Added by Michael Walker on October 10, 2012 at 10:52am — No Comments

Big Data Analytics Maturity Model

Phase / Impact	The Old World			The New Era
	Pilot	Departmental Analytics	Enterprise Analytics	Big Data Analytics
Staff Skills (IT)	Little or no expertise in analytics – basic of knowledge BI tools	Data warehouse team focused on performance, availability and security	Advanced data modelers and stewards key part of the IT department	Business Analytics Competency Center (BACC) that includes 'data scientists'
Staff Skills (Business/IT)	Functional knowledge for BI tools	Few business analysts – limited usage of advanced analytics	Savvy analytical modelers and statisticians utilized	Complex problem solving integrated into Business Analytics Competency Center (BACC)
Technology & Tools	Simple historical BI reporting and dashboards	Data warehouse implemented, broad usage of BI tools, limited analytical data marts	In database mining, usage of high performance computing & analytical appliance	Widespread adoption of appliances for multiple workloads. Architecture and governance for emerging technologies
Financial Impact	No substantial financial impact. No ROI Models in place	Certain revenue generating KPI's in place with ROI clearly understood	Significant revenue impact (measured and monitored on a regular basis)	Business strategy & competitive differentiation is based on analytics
Data Governance	Little or none (Skunk works)	Initial data warehouse model and architecture	Data definitions & models standardized	Clear master data management strategy
Line of Business	Frustrated	Visible	Aligned (including LoB executives)	Cross-departmental (with CEO visibility)
CIO Engagement	Hidden	Limited	Involved	Transformative

Source: IDC Asia/Pacific Business Analytics Practice (July, 2011)

See: …

Continue

Added by Michael Walker on October 3, 2012 at 10:18am — No Comments

Data Fundamentals

Becoming a data and evidence driven organization provides significant competitive advantage. Speed and accuracy of insight, delivered across any device including smart phones and tablets, means organizations can make better,...
Continue

Added by Michael Walker on September 26, 2012 at 10:00am — No Comments

Predictive, Descriptive, Prescriptive Analytics

Descriptive, Prescriptive, Predictive Analytics Competency

The goal of Data Analytics (big and small) is to get actionable insights resulting in smarter decisions and better business

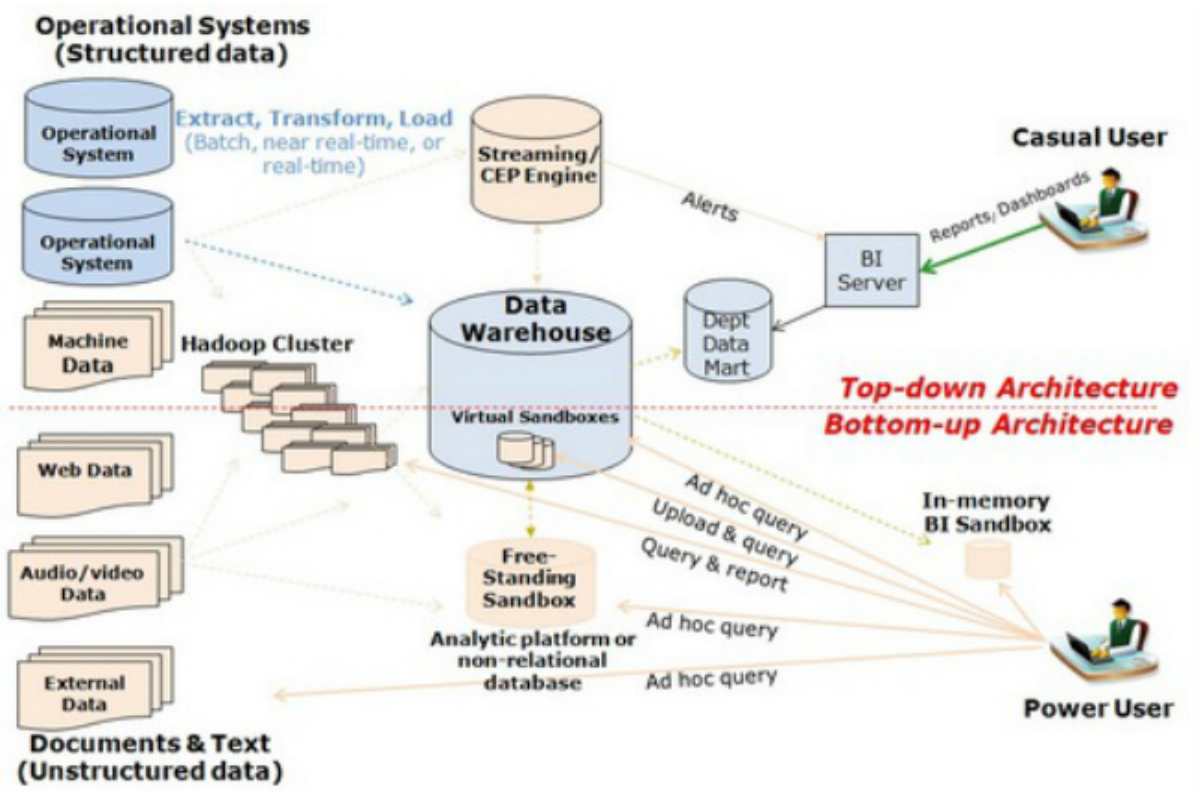
outcomes. How you architect business technologies and design data analytics processes to get valuable, actionable insights varies.

It is critical to…

Continue

Added by Michael Walker on September 19, 2012 at 11:57am — No Comments

Modern BI Architecture & Analytical Ecosystems



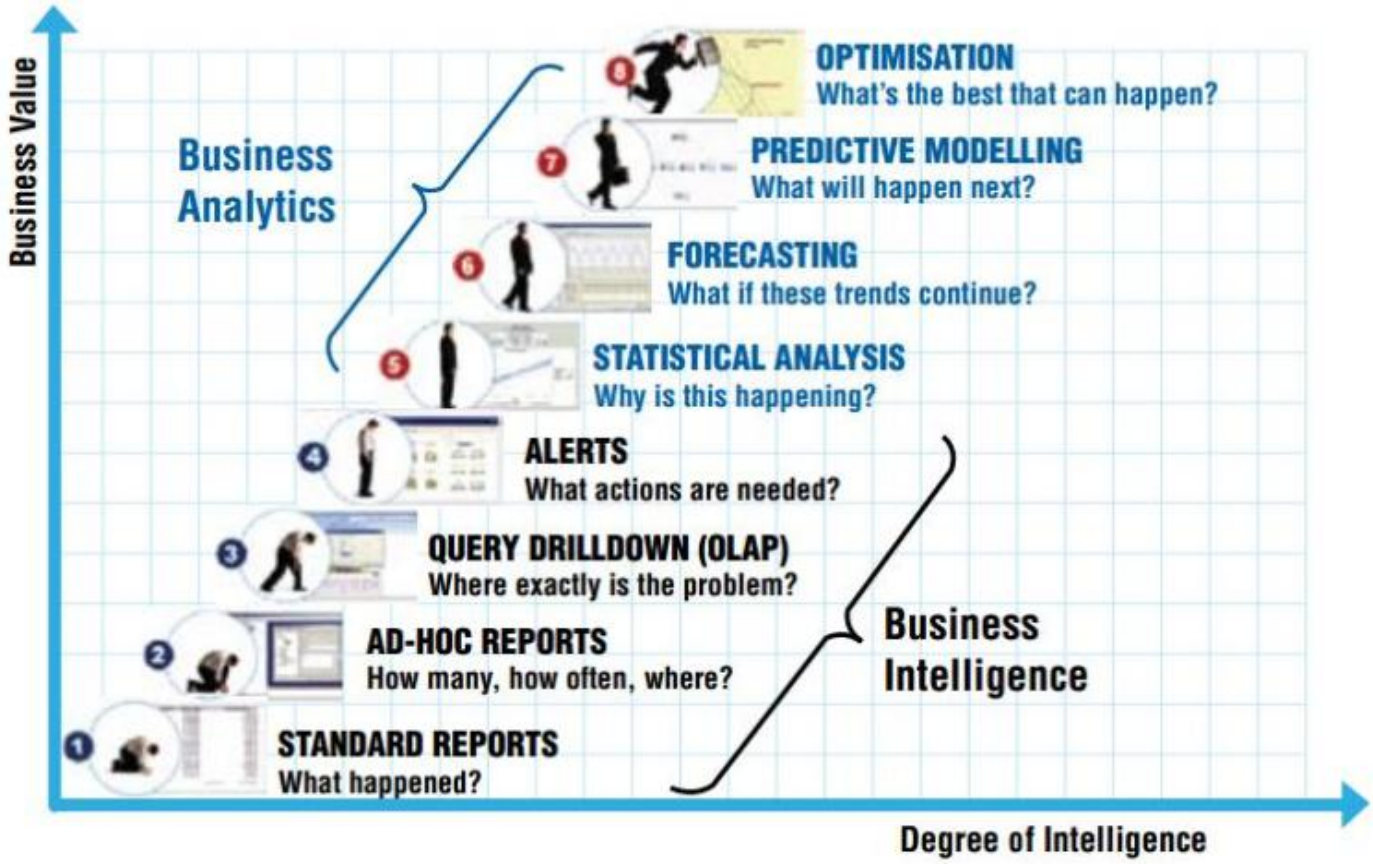
The goal is to design and build a data warehouse / business intelligence (BI) architecture that provides a flexible, multi-faceted analytical ecosystem for each unique organization.

A traditional BI architecture has analytical processing first pass…

Continue

Added by Michael Walker on September 12, 2012 at 11:53am — No Comments

Eight Levels of Analytics for Competitive Advantage



…

Continue

Added by Michael Walker on September 6, 2012 at 11:30am — 1 Comment

Big Data Vendor Landscape

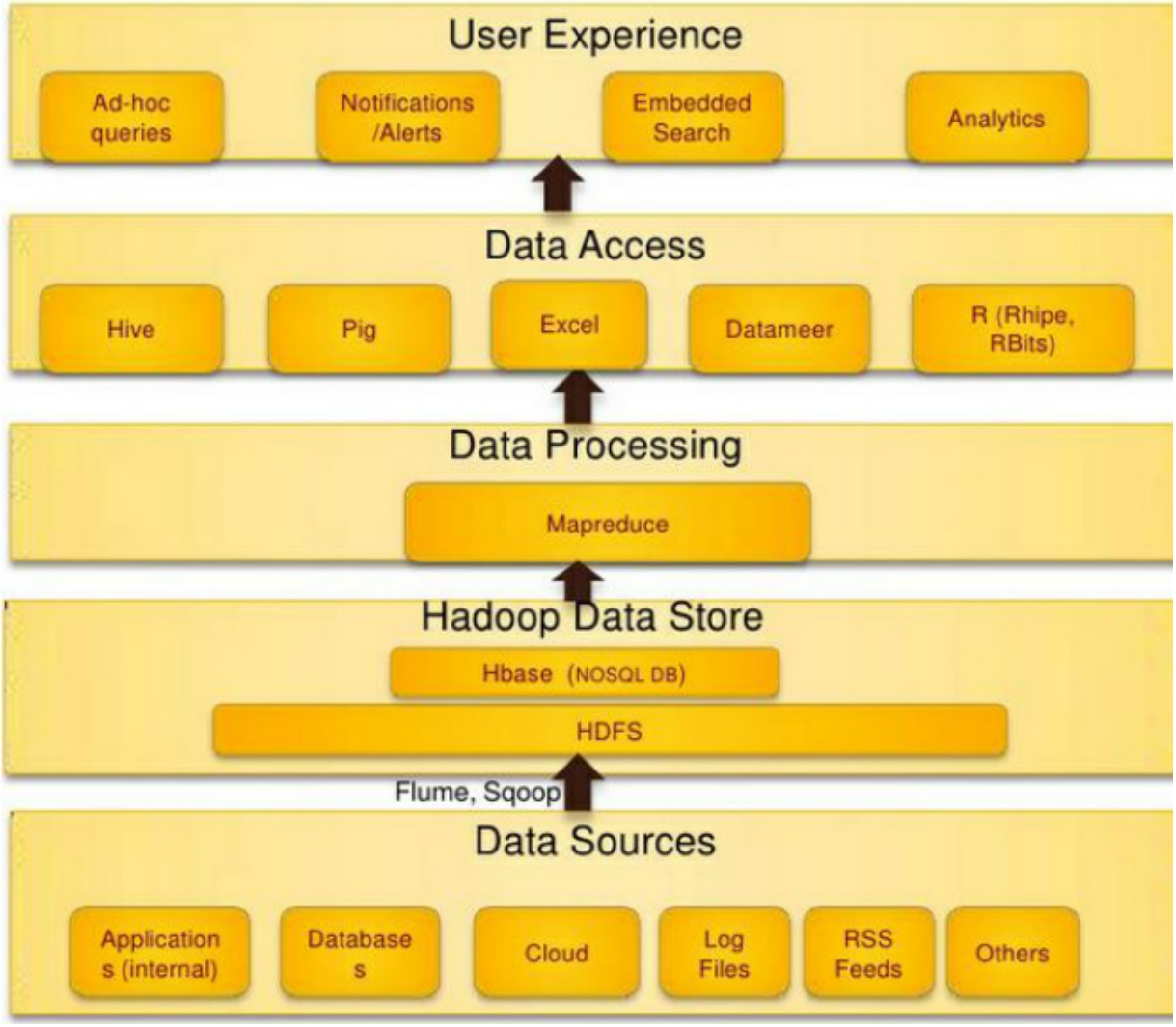
Big Data Vendor Landscape…

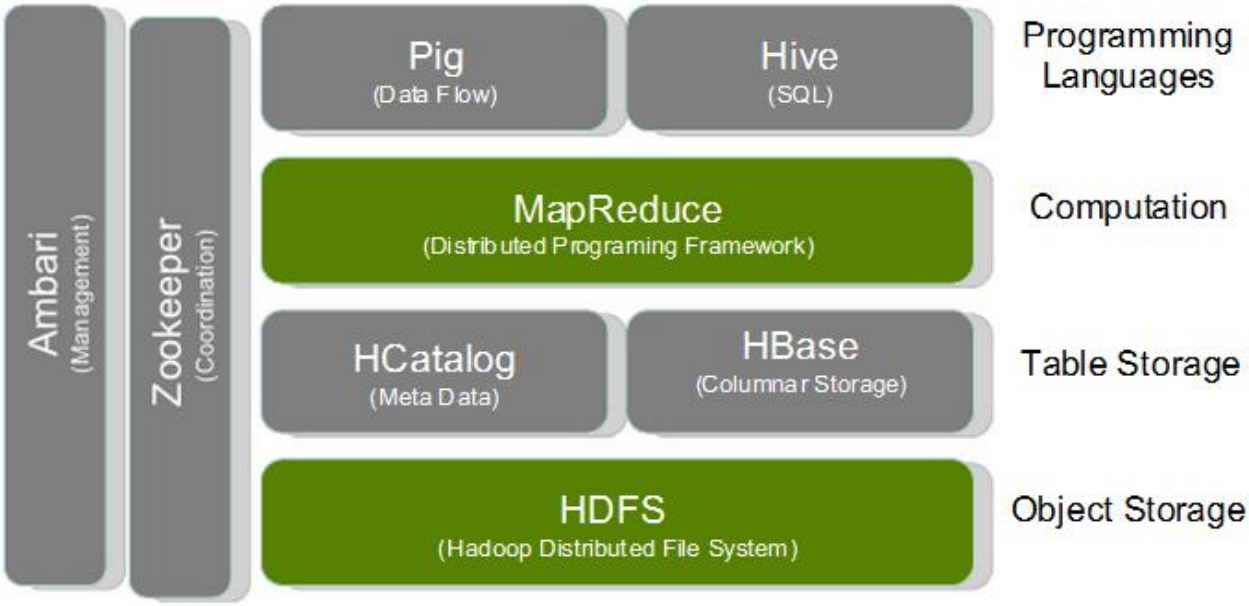
Hardware	Big Data Distributions	Data Management Components	Analytics and Visualizations	Services
• Storage • Servers • Networking Vendors include Dell, HP, IBM, Cisco.	• Community Hadoop distributions • Enterprise Hadoop distributions • Non-Hadoop Big Data frameworks Vendors include Cloudera, IBM, MapR, LexisNexis, Microsoft.	• NoSQL databases • Data integration • Data quality and governance Vendors include DataStax, IBM, Informatica, Syncsort.	• Analytic development platforms • Advanced analytics applications • Data visualization tools • Business intelligence applications Vendors include Karmasphere, Tresata, Datameer, SAS Institute, Tableau, Revolution Analytics.	• Consulting • Training • Software maintenance • Hardware maintenance • Hosting/cloud Vendors include Think Big Analytics, Amazon Web Services, Accenture, as well as services associated with enterprise distributions (e.g. Cloudera).
Next Generation Data Warehouse • MPP, columnar data warehouse appliances. • In-memory analytics engines Vendors include EMC Greenplum, HP Vertica, Teradata Aster Data, IBM Netezza, SAP, Microsoft, Kognitio.				

Continue

Added by Michael Walker on August 30, 2012 at 2:58pm — No Comments

Hadoop Technology Stack





The Hadoop stack includes more than a dozen components, or subprojects, that are complex to deploy and manage.…

Continue

Added by Michael Walker on August 22, 2012 at 9:40am — No Comments

Stop Following – Don't email me when this member adds new blog posts

Latest Blog Posts

- Untestable & Unreasonable Assumptions in Models is Data Science Malpractice
- Data Diversity and Integration Trumps Big Data for Better Decision-making
- Data Science for Improved Democracy
- Data Science is Dead - Long Live the Data Scientist
- Caveat Data Scientist: Public Trust Low for Science
- Big Data is Stupid Data
- Big Data Technology Vendor Consolidation

Most Popular Blog Posts

- Batch vs. Real Time Data Processing
- Data Scientists vs. Data Engineers
- Data Science Summer Reading List 2013
- Predicting the Super Bowl
- Random Forests Algorithm
- Untestable & Unreasonable Assumptions in Models is Data Science Malpractice
- Selecting the Right Business Intelligence and Analytics Platform

Blog Topics by Tags

- Data (63)
- Science (28)
- Analytics (25)
- Big (17)
- Business (16)
- Intelligence (15)
- of (10)
- Predictive (9)
- Hadoop (8)
- Scientists (6)
- Models (6)
- Bias (5)
- Professional (5)
- Machine (4)
- BI (4)
- Conduct (4)
- learning (4)
- Prescriptive (4)
- Algorithms (4)
- Descriptive (4)
- Code (4)
- Association (3)
- Platform (3)

- Batch (3)
- Markov (3)
- Vendor (3)
- Ethics (3)
- 2014 (3)
- machine (3)
- Scientist (3)
- Internet (3)
- Security (3)
- Processing (3)
- Algorithm (3)
- Team (3)
- algorithms (3)
- Learning (3)
- Signal (3)
- Things (3)
- NoSQL (3)
- Noise (3)
- Reading (2)
- carlo (2)
- Dr. (2)
- Malpractice (2)
- regression (2)
- Assumptions (2)
- data (2)
- Management (2)
- decision (2)
- MapReduce (2)
- Ecosystems (2)
- Hidden (2)
- Complex (2)
- Return (2)
- Stack (2)
- Variety (2)
- Confirmation (2)
- Engineers (2)
- simulations (2)
- Meaning (2)
- Veracity (2)
- intelligence (2)
- Information (2)
- Evidence (2)
- trees (2)
- Infrastructure (2)
- Correlation (2)
- Strategy (2)
- Scientific (2)
- Experiments (2)
- monte (2)
- Processes (2)
- Investment (2)
- Database (2)
- Ioannidis (2)
- Selection (2)
- Profession (2)
- John (2)
- Service (2)
- Causation (2)
- Forecasting (2)

- Language (2)
- on (2)
- Python (2)
- Chain (2)
- Law (2)
- Chains (2)
- Analytical (2)
- R (2)
- ROI (2)
- List (2)
- Apache (2)
- bayesian (2)
- probability (2)
- Architecture (2)
- Cognitive (2)
- Asset (1)
- Twitter (1)
- Haboob (1)

Monthly Archives

2014

- December (1)
- November (1)
- October (1)
- September (3)
- August (1)
- July (2)
- June (1)
- May (1)
- April (1)
- March (2)
- February (2)
- January (1)

2013

- December (1)
- November (1)
- October (1)
- September (3)
- August (1)
- July (2)
- June (1)
- May (2)
- April (4)
- March (2)
- February (2)
- January (2)

2012

- December (2)
- November (2)
- October (4)
- September (1)
- August (2)

Welcome to
Data Science Central

Sign Up
or Sign In

Or sign in with:

-
-
-

Follow Us

@DataScienceCtrl | RSS Feeds

Top Content

1 **Data science without statistics is possible, even desirable**

2 **The Only Skill you Should be Concerned With**

3 **Data Scientist: Owning Up to the Title**

4 **Implementing a Distributed Deep Learning Network over Spark**

5 **Popular Software Skills in Data Science Job postings.**

6 **5 basic rules of data organization**

7 **63 Machine Learning, Data Science, Big Data Resources and Articles**

8 **Data Science + Behavioral Science = Force Multiplier ! 3 User Stories**

9 **4 easy steps to becoming a data scientist**

10 

Trends in Big Data Vs Hadoop Vs Business Intelligence

- RSS
- View All

Resources

38 Seminal Articles Every Data Scientist Should Read

Black-box Confidence Intervals: Excel and Perl Implementation

Data Science Cheat Sheet

16 analytic disciplines compared to data science

10 types of regressions. Which one to use?

Selection of best articles from our past weekly digests

Best kept secret about data science competitions

Free stuff for publishers, authors, bloggers, professors, event organizers, companies etc.

Must Read Before Attending Any Data Science Job Interview

Big Data Poster

Videos

- 

  DSC Webinar Series: How United Way Embraces Data Visualization To Drive Social Impact

  Added by Tim Madison 0 Comments 0 Likes

- 

  DSC Webinar Series: Data Transformation and Acquisition Techniques, to Handle Petabytes of Data

  Added by Tim Madison 0 Comments 0 Likes

- Add Videos

- View All

Announcements

Big Data, Hadoop, Data Science Courses: Big Discount Ends 12/22

Using visualization to share the human impact of numbers - Whitepaper

Creating a Global Analytics Culture - San Diego, February 12-13

Tips for making Hadoop a productive environment for Data Scientists

Top Data Science Trends for 2015 - Pivotal Webinar

The 2014 Data Science Salary Report

Visual Best Practices: A Guidebook

Engaging employees in data and transforming your business - Webinar

Apache Hadoop and Data Science Summits - San Diego, February 12-13

Why Most Big Data Projects Fail - White Paper