



Big Analytics: *A Next Generation Roadmap*

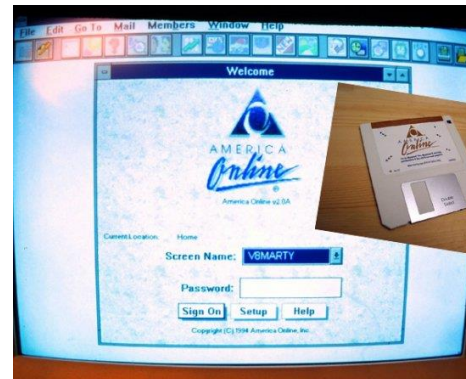
Cloud Developers Summit & Expo: October 1, 2014

Neil Fox, CTO: **SoftServe, Inc.**

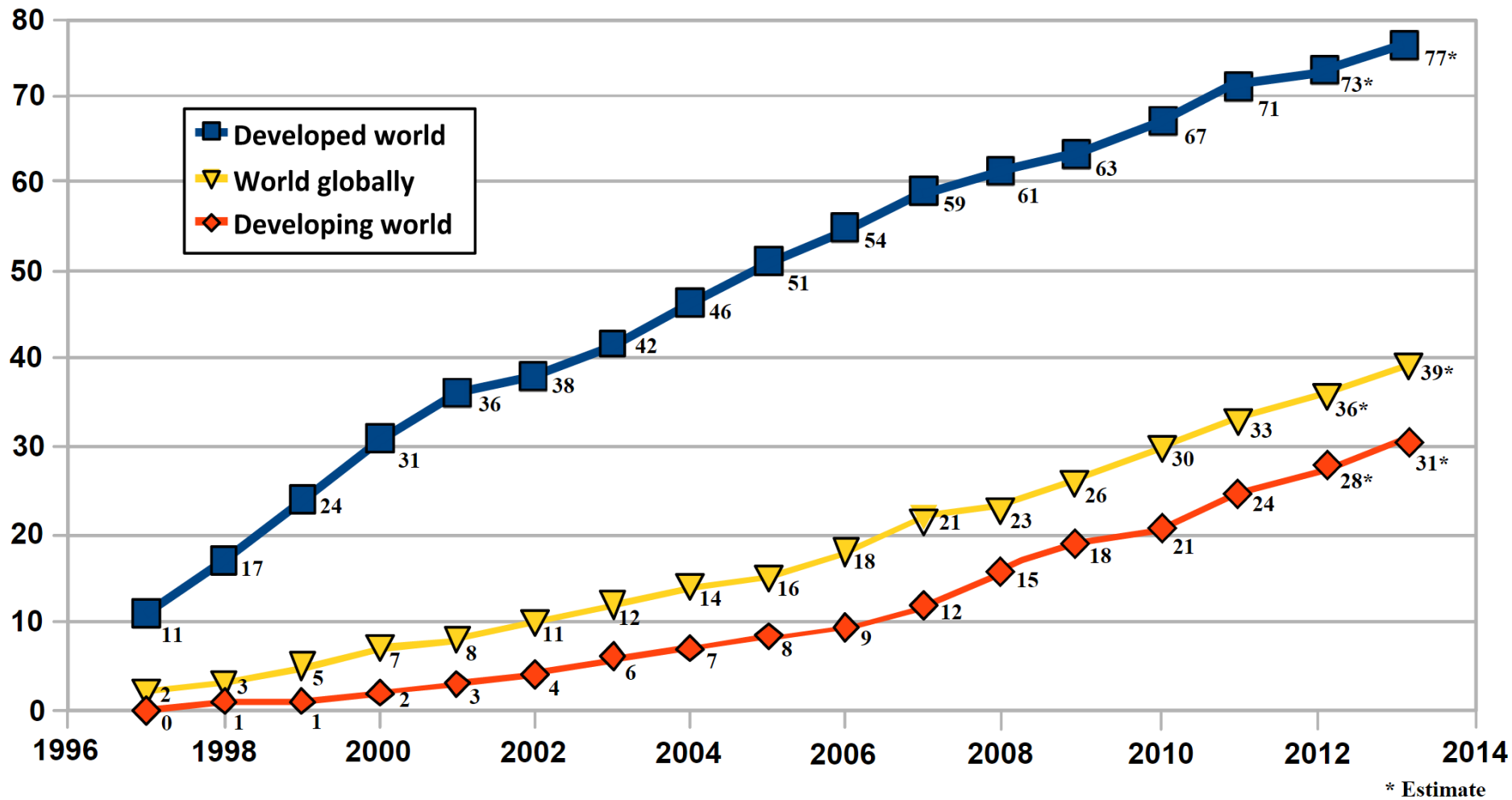
© 2014 SoftServe, Inc.

Remember Life Before The Web?

1994



Even Revolutions Take Time



* Estimate

Big Analytics?

**BIG
DATA**

**INTEGRATED
ANALYTICS**

**BIG
ANALYTICS**

Big Analytics also describes the technology solutions used by giants such as Google®, Amazon®, and Facebook® to process enormous data to provide you the best in internet, real-time services.



Value Proposition

Public Safety:

Durham, NC Police use Predictive Analytics to reduce crime rate by 50%

Fraud Prevention:

Insurance company uses Predictive Analytics to save \$12 m annually

Supply Chain Optimization

Retailers using Predictive Analytics to forecast product demand, price, promotion and inventory management

Healthcare

Impacting every aspect of the healthcare system – giving more personalized data to patients, providers and payers. Areas of focus include lifestyle, diet, exercise, research, and clinical trials

Retail:

Macy's boosts same store sales by 10% using 10s of millions of data points from twitter, social media, in-store and on-line.

...And just how powerful is this stuff?

Marketing (e.g. Target):

".... computers crawled through the data, he was able to identify about 25 products that, when analyzed together, allowed him to assign each shopper a “pregnancy prediction” score. More important, he could also estimate her due date to within a small window, so Target could send coupons timed to very specific stages of her pregnancy." [1]

Government (via SAP survey):

"87% of federal and 75% of state IT officials believe that real-time Big Data has the potential to save a significant number of lives." [2]

[1] <http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=6&r=2&hp&>

[2] http://www.techamericafoundation.org/content/wp-content/uploads/2013/02/SAP_INFOGRAPHIC_BIG-DATA_Final11.pdf

A Success Story

The Durkheim Project



- *Developed linguistics-driven prediction models to estimate the risk of suicide. & Reached 70% accuracy in cohort distinction [3]*
- *Deployed a real-time big data framework to capture opt-in mobile and social media data from veterans*
- *Fast Company said “This may be the most vital use of big data we’ve ever seen.” [4]*



[3] www.durkheimproject.org

[4] <http://www.fastcolabs.com/3014191/this-may-be-the-most-vital-use-of-big-data-weve-ever-seen>

So What is Slowing Adoption Down? ...

1. Resistance to change

“... there's the drag exerted by relational database administrators who badly want to stick to what they know.” [5]

2. Volume, Variety, and Velocity

“... big data problems have just as much to do with changing how you do data querying and processing as they do with handling the oft-cited "three V's" -- the big data parameters of volume, variety, and velocity.”

- *Volume (of data under management) - Data is growing from the Terabytes to the Petabytes... for everyone*
- *Velocity (of transactions)- NoSQL simply lets you access your data differently*
- *Variety (of data) – Unstructured?, structured?, semi-structured?*

[5] <http://www.informationweek.com/big-data/software-platforms/big-data-how-to-pick-your-platform/d/d-id/1315609?ngAction=register>

Big Analytics Engineering Challenges

How to achieve **Low Latency** for personalized customer experience in real-time?



Consumers



Intelligent Agents

Real Time Intelligence



How to implement **Self-Service** with high **Data Quality** over terabytes and petabytes?



Business Users

Business Reporting



How to improve **System Performance** for Data Science/ Analytics team?



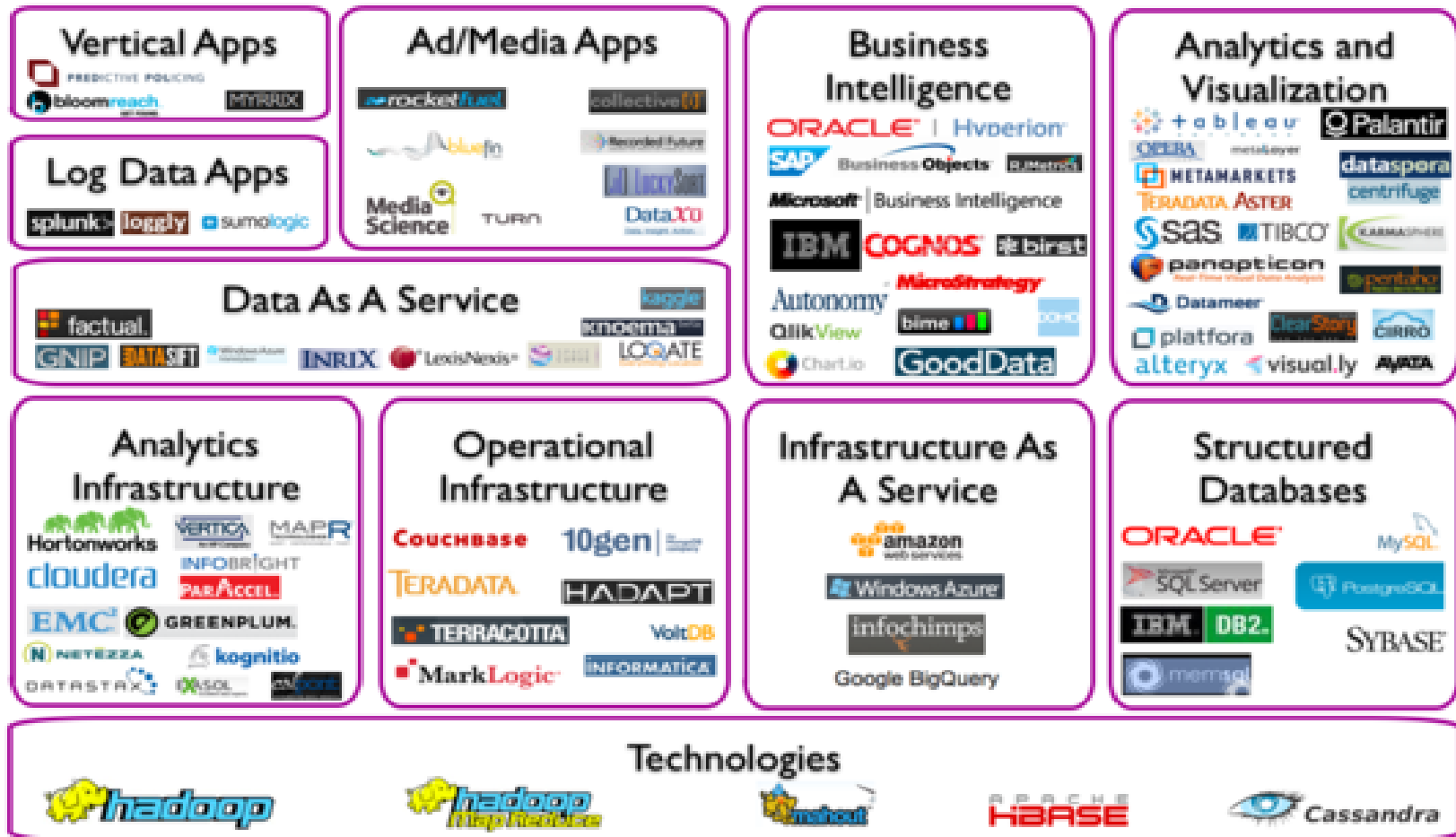
Data Scientists/
Analysts

Data Discovery



Complex Environment

Big Data Landscape



Copyright © 2012 Dave Feinleib

dave@vcdave.com

blogs.forbes.com/davefeinleib

Sample Technologies



But now Hadoop...

Hadoop In Action

How is Hadoop used, or how will it be used, at your organization?

Running analytics



Business intelligence



For extract, transform, and load functions



Archiving



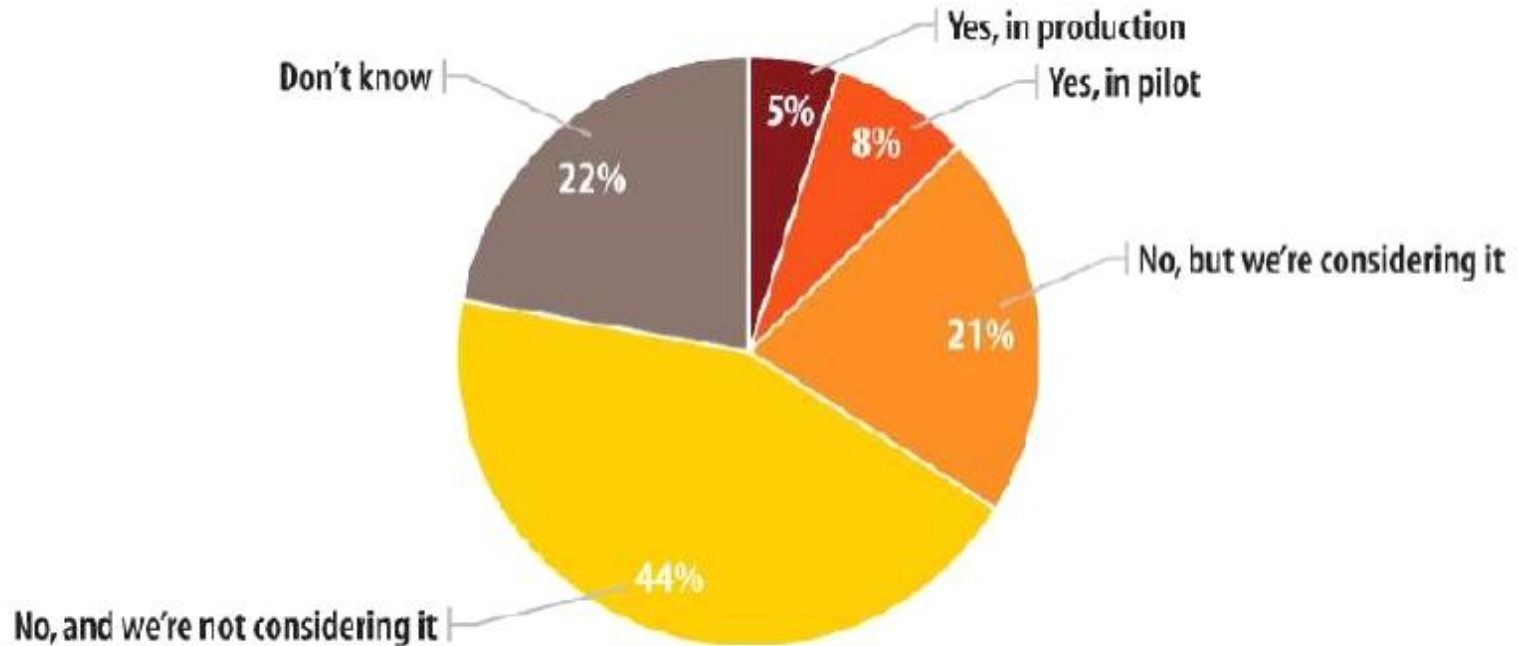
Data: InformationWeek 2014 State of 321 Database Technology Survey of 956 business technology professionals using or considering using Hadoop, January 2014

R7770314/12

InformationWeek CONNECTING THE BUSINESS TECHNOLOGY COMMUNITY

...though the adoption is still behind

Does Your Organization Run Hadoop?



Data: InformationWeek 2014 State of Database Technology Survey of 956 business technology professionals, January 2014

R7770314/11

InformationWeek CONNECTING THE BUSINESS TECHNOLOGY COMMUNITY


Big Analytics Reference Architectures and Practices

- ROI-driven Big Analytics systems design based on proven Architectures and Technologies
- Maximum efficiency with the lowest Cost per Terabyte

Spark SQL

/Analytics/Query & Reporting/Distributed Query Processor

Description: Based on Spark—an in-memory distributed computing engine (alternative to MapReduce)—allows running SQL and HiveQL on Spark. Spark SQL is an ancestor of Shark.




Consequences:

- ★★ Query capabilities — based on SQL, supporting most of HiveQL features in memory
- ★★ Performance — considered as one of the fastest at the moment, significantly faster than Hive
- ★ Compatibility — for now supports only JDBC, can work through ODBC/DDBC
- ★★★ Reliability — supports long-run recovery
- ★ Maturity — currently in alpha stage, that counts its history since 2011

Apache Storm

/Analytics/Event Stream Processor

Description: Storm is a distributed real-time computation system. Similar to how Hadoop provides batch processing, Storm provides for doing real-time computation.




Consequences:

- ★★★ Real-time capabilities — one of the lowest latency, supports both one-at-a-time and batch processing
- ★★ Analytics capabilities — event aggregation, distributed RPC
- ★ Compatibility — integrates with other systems such as RabbitMQ/AMQP, Kafka
- ★★★ Reliability — reassigns tasks in case of failure and exactly-once delivery guarantee
- ★★★ Maturity — the initial release was in 2010, an impressive number of adopters

Elasticsearch

/Analytics/Distributed Search Engine

Description: An open source search engine based on Lucene. It provides distributed search capabilities with a RESTful API. It stores data in free JSON documents. Along with Logstash and Kibana, it forms the ELK stack to collect, index and visualize data.



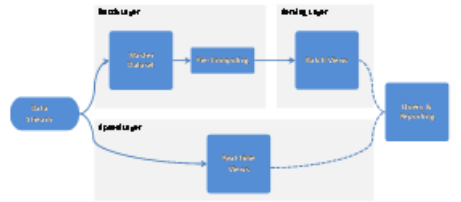
Consequences:

- ★★ Analytics capabilities — provide advanced facets such as geo, has_child queries
- ★★ Real-time capabilities — enable near real-time search
- ★★★ Reliability — implements sharding, replication and recovery
- ★ Maturity — the first version released in 2010, Hortonworks and MapR
- ★★★ License cost economy — released under the Apache License

Lambda Architecture (Hybrid)

/Reference Architecture/Data Analytics

Description: This reference architecture enables real-time operational and historical analytics in the same duration. While the batch layer is based on non-relational techniques (usually Hadoop), the speed layer is based on streaming techniques to support strict real-time analytics requirements.



Consequences:

- ★★★ Scalability — can scale keeping petabytes
- ★★★ Real-time analysis — streaming approach provides extremely low data latency
- ★★★ Extensibility — can be easily extended with new data formats and sources
- ★★ Ad-hoc analysis — ad-hoc real-time query support is more difficult than in relational architecture

Sample implementations: Real-time Intelligence, Data Discovery, Business Reporting

Data Refinery (Hybrid)

/Reference Architecture/Data Analytics

Description: This reference architecture does not rely on a single data source. It uses NoSQLs for storage, and a distributed processing engine for real-time analytics.

Non-relational

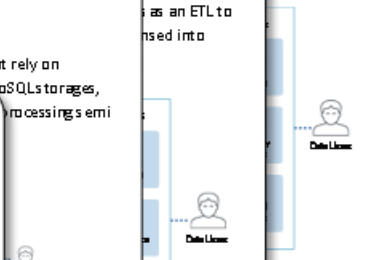
/Reference Architecture/Data Analytics

Description: This reference architecture does not rely on a single data source. It uses NoSQLs for storage, and a distributed processing engine for real-time analytics.

Extended Relational

/Reference Architecture/Data Analytics

Description: This reference architecture does not rely on a single data source. It uses NoSQLs for storage, and a distributed processing engine for real-time analytics.



Consequences:

- ★★★ Scalability — can scale keeping petabytes
- ★★★ Real-time analysis — streaming approach provides extremely low data latency
- ★★★ Extensibility — can be easily extended with new data formats and sources
- ★★ Ad-hoc analysis — ad-hoc real-time query support is more difficult than in relational architecture

So What to Do?

To Properly Frame a Big Data + Analytics Project

1) What is the business goal?

2) What data can we get in support of this?

3) How can we display business intelligence intuitively?

4) Iterate, iterate, iterate

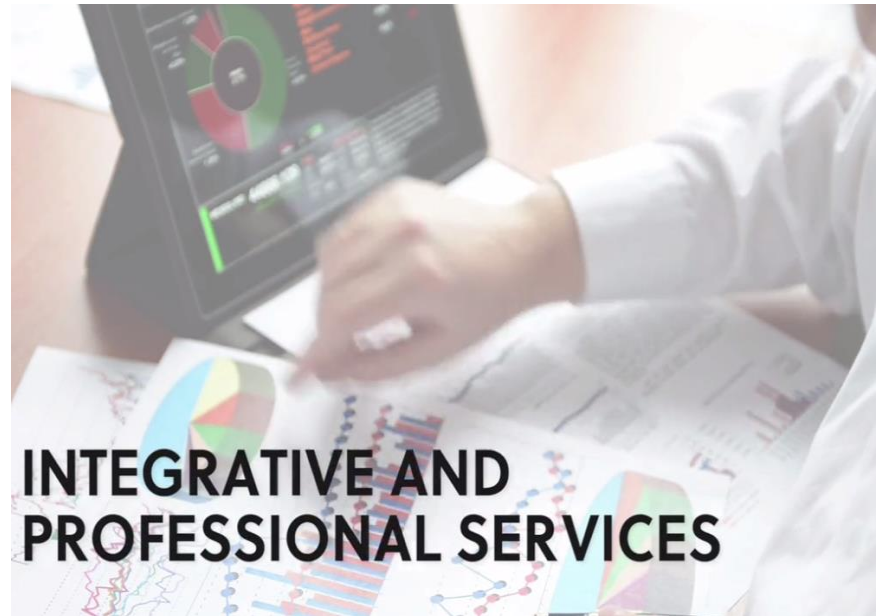
“Start with a focused, business-driven project, make sure the data is consistent with your vision and then apply advanced analytics without moving beyond human-understandable decisions.” [6]

[6] <http://www.informationweek.com/big-data/big-data-analytics/big-data-success-3-companies-share-secrets/d/d-id/1111815?>

SoftServe, Inc. is a unique software development partner, offering specialized outsource technical staffing... and integrative, and professional services to some of the world's premier technology companies.



**SPECIALIZED
OUTSOURCED
TECHNICAL
STAFFING**



**INTEGRATIVE AND
PROFESSIONAL SERVICES**

Our Unique Integrated Approach

Abiliton™

Abiliton Big Analytics

SoftServe's adaptive best practice framework for Big Data/Business Analytics transformation and optimization

People

Competence Development

- Knowledge Model
- Performance Management Practices
- Training Catalog

Organization Structure

- Optimal Team Structure
- Roles and Balance

Process

SDLC Optimization

- Governance
- Project Management
- Business Analysis
- Software Engineering
- Quality Control
- DevOps
- Metrics for Continuous Improvement
- Project Status Dashboard

Technology



Data Science & Analysis

- Predictive Modeling
- Statistical Analysis
- Standard and Ad-Hoc Reports

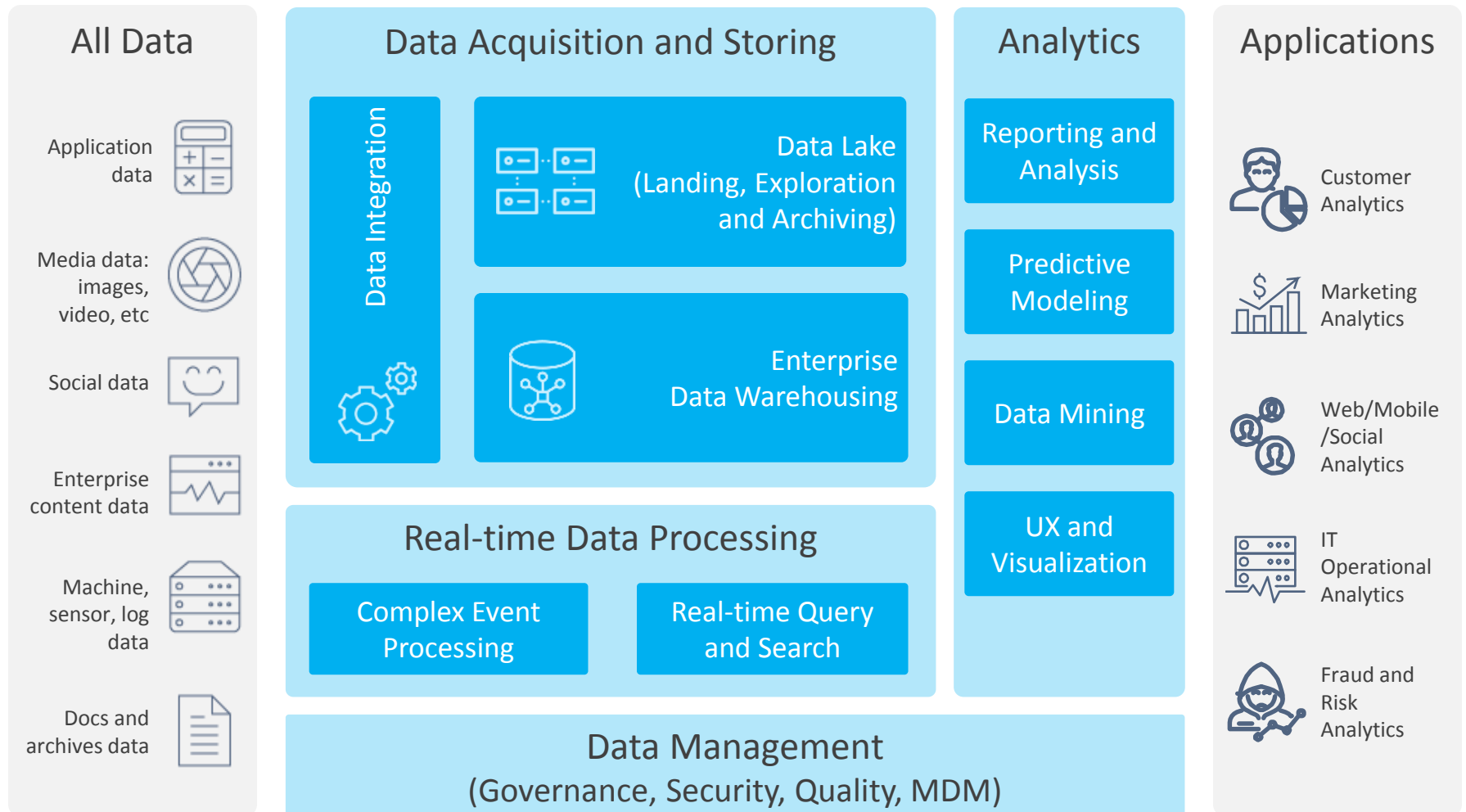


Data Engineering

- Solution Architecture
- System Modernization
- System Optimization
- Operations Automation

Big Analytics: Technology Framework

A modern integrated approach for solving Big Data/Business Analytics needs across multiple verticals and domains



Big Analytics Case Study: Network Security

Business Goals:

1. Provide reporting platform in the cloud for services & applications usage analysis
Charge customers based on the platform they are using, number of consumers' applications etc.

Technical Specs:

Machine generated data

Big Data: 7.5BLN log records per day

Near real-time reporting

Reports which "touch" billions of rows

Solution:

ETL - Talend

DW - HP Vertica/ InfoBright EE

OLAP - Pentaho

BI - JasperServer Pro



Big Analytics Case Study: Online Analytics

Business Goals:

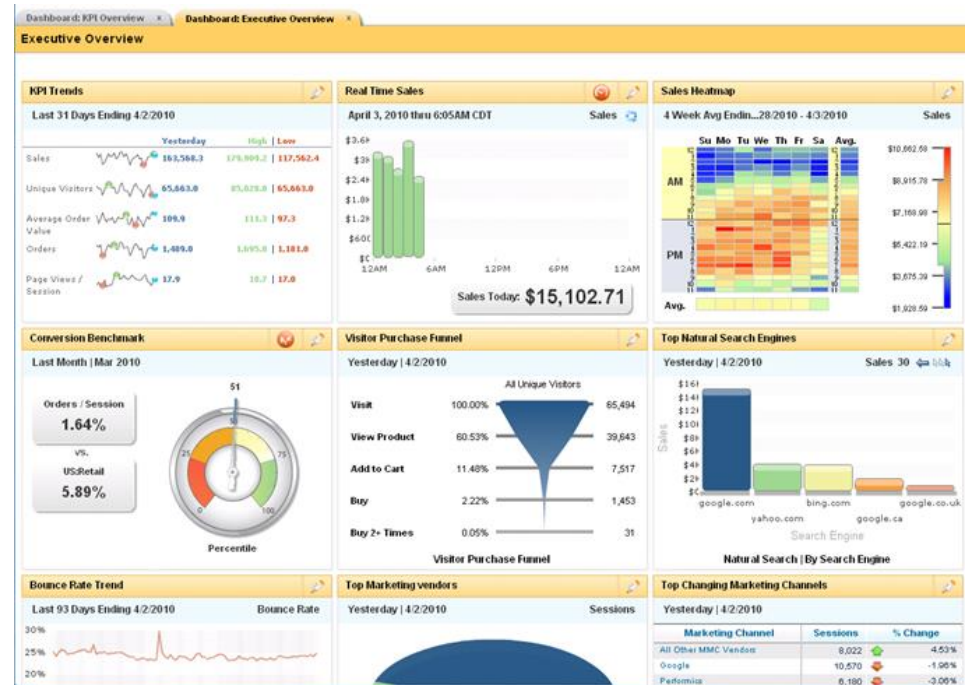
Insights and optimization of all web, mobile, and social channels
Optimization of recommendations for each visitor
High return on online marketing investments

Technical specs:

Big Data > 1PB
10+ GB per customer/day
10+ Hadoop Clusters
15+ Aster Data Clusters

Solution:

Hadoop/HBase/Hive
Aster Data
Oracle
Java/Flex



Takeaway: Metrics for Success

- ***Qualitative Performance: For example, compelling visualizations that make tasks easier (i.e. less complex)***
- ***Quantitative Performance: For example, maximizing systems ROI, reducing TCO, or even saving lives...***
- ***Compliance and Data Governance: For example, privacy concerns & jurisdictional issues***

Remember Life Before Big Analytics?

2025?

- Applications were generic
- Doctors (and patients) had one data point per year
- Doctors had to rely on their own ability to research
- Shoppers had to search for deals
- Companies had exabytes of data they were not using



*Empowering your Business
through Software Development*



Thank you!

US OFFICES

Austin, TX
Fort Myers, FL
Boston, MA
Newport Beach, CA
Salt Lake City, UT

EUROPE OFFICES

United Kingdom
Germany
Netherlands
Ukraine
Bulgaria

EMAIL

info@softserveinc.com

WEBSITE:

www.softserveinc.com

USA TELEPHONE

Toll-Free: 866.687.3588
Office: 239.690.3111