Cloud Analytics Playbook
Booz | Allen | Hamilton
An enormous amount of valuable information is out there, waiting to be transformed into differentiating services. Booz Allen Hamilton uses its Cloud Analytics Reference Architecture to build technology infrastructures that can withstand the weight of massive datasets—and deliver the deep insights organizations need to drive innovation.

PREFER TO READ THIS ON YOUR IPAD?

Search “Booz Allen” at the iTunes App Store or simply scan the QR code.
A majority of executives believe their companies are unprepared to leverage their data. We look at why that is and how to change it.
Extracting True Insights

THE GROWING DATA ANALYSIS GAP

We are living in the greatest age of information discovery the world has ever known.

According to recent industry research, we now generate more data every 2 days than we did from the dawn of early civilization through the year 2003 combined. And data rates are still growing—approximately 40% each year.

Fueled in large part by the more than five billion mobile phones in use around the globe, our world is increasingly measured, instrumented, monitored, and automated in ways that generate incredible amounts of rich and complex data. Unfortunately, the number of big data analysts and the capabilities of traditional tools aren’t keeping pace with this unprecedented data growth.

At Booz Allen, we’ve watched this trend for some time now—we call it the “data analysis gap.” It’s clear that data has outstripped common analytics tools and staffing levels. In order to move forward, organizations must be able to analyze data on a massive scale and quickly use it to provide deeper insights, create new products, and differentiate their services.

By 2020, the amount of information in our economy will grow 44 times. Very few organizations are prepared for this wave of data.

(Source: IDC)
Preparing for
What’s Ahead

The ability to compete and win in the information economy will come from powerful analytics that draw insights and value from data, and from high-fidelity visualizations that present those insights in impactful, intuitive ways. Both will become key influencers of corporate decision making and consumer purchasing.

Many of the world’s IT systems are not ready for the technology revolution happening as organizations seek to transform how they use data. Their infrastructures face three major challenges:

- **Volume**: Not enough storage capacity and analytical capabilities to handle massive volumes of data
- **Variety**: Data comes in many different formats, which can be difficult and expensive to integrate
- **Velocity**: Inability to process data in real time in order to extract the most value from it

To help organizations overcome these hurdles and prepare for what’s next, Booz Allen has pioneered strategies for the implementation of the Digital Enterprise—a way of using technology, machine-based analytics, and human-powered analysis to create competitive and mission advantage.

A Framework for the Future

Booz Allen has a framework for intelligently integrating cloud computing technology and advanced analytic capabilities, called the Cloud Analytics Reference Architecture. The Architecture is designed to solve compute-intensive problems that were previously out of reach for most organizations, including large-scale image processing, sensor data correlation, social network analysis, encryption/decryption, data mining, simulations, and pattern recognition.

At the core of the Architecture are systems that accommodate petabytes of data at reasonable cost and allow analytics to run at previously unattainable scales in reasonable amounts of time. However, human insights and action are still the fundamental drivers.

The purpose of the Architecture is to allow machines to do 80% of the work—the mundane tasks they are best suited for—and enable people to do the 20% of the work they do best, tasks that involve analysis and creativity.
The Transformative Power of Cloud Analytics

Booz Allen is the leader in the emerging field of cloud analytics. Our unique approach combines cloud and other technologies with superior analytic tradecraft to create breakthroughs in how organizations capture, store, correlate, pre-compute, and extract value from large sets of data.

Analysis using standard cloud computing solutions extends basic analytic techniques to large or very large datasets. This is a logical entry point for cloud solutions because cloud technology is the most efficient, cost-effective way to run analytics on large amounts of data.

To understand the power of cloud analytics, it helps to see the progression from basic data analytics performed in most organizations today. As an infrastructure is built out along the continuum to cloud analytics, the size and scale of data it can process increases along with the ability to drive performance and improve decision making.

Advanced analytics is where predictive capabilities are brought into the mix. It’s generally used to evaluate the future impact of strategic decisions. However, it represents a step back in terms of the size of datasets that can be manipulated.

Cloud analytics transcends the limits of the other forms of analysis. It delivers insights to answer previously unanswerable questions such as:

- How can we gain competitive advantage in our market space?
- Where can we save money within our organization?
- How should we turn our data into a product?
Booz Allen clients have a wide variety of data analysis challenges and IT infrastructures. Our flexible, scalable Cloud Analytics Reference Architecture has three stages or entry points to accommodate these differences.

In each stage, we enable shifts in technology investments while helping manage risk and maximize the reward. That means leveraging the assets you already own and taking logical steps to add what’s needed. This is the only way to build a structure to instrument data so you can truly experience breakthrough analytics.

**STAGE 1**

The focus is on saving money and reducing risk. You may have already begun some of these initiatives; we leverage what’s working now as we discover new ways to increase efficiency.

**STAGE 2**

We begin to modernize applications to handle the demands of advanced analytics. Faster, reusable, and more intuitive applications will enable everyone in your organization to work smarter.

**STAGE 3**

Significant improvements in performance are realized when you achieve success in managing the flow of information at scale and derive the fullest value from your data.

**IT EFFICIENCIES**
- Data center consolidation
- Server and data consolidation
- Increased automation
- Modernized security posture and metrics
- Reduced licensing costs

**SMART DATA**
- Enhanced Enterprise Data Architecture
- Clarify pedigree (data tagging)
- Multidimensional indexing
- Adopt distributed database
- Reusable applications

**CLOUD ANALYTICS**
- Create deep insight into relevant mission data at scale
- Ask and answer previously unanswerable questions
A Better Approach

As the leaders in cloud analytics, Booz Allen has a proven approach delivered by some of the industry’s best talent. Here’s why we’re different:

Technical framework
Our Architecture combines the collective experience of thousands of people who have road tested technologies from across the cloud solution landscape in hundreds of client organizations, ranging from the U.S. Federal Government to commercial and international clients.

Best practices
We have an exclusive set of lessons learned and breadth of technical knowledge that saves time and money while reducing risk.

Core principles
These are “rules of the road” we’ve developed to build the most effective solution with the highest return on investment. They encompass everything from how data should be stored to how to improve relationships with the end users of your data.

Critical skill sets
We bring technologists as architecture and solutions specialists, domain experts who know your industry and your data, and data scientists who explore and examine data from disparate sources and recommend how best to use it. No one else in the industry offers a better combination of talent.

Vendor neutrality
Our approach utilizes a broad ecosystem of products and custom systems culled from an exhaustive survey of available options. In the crowded, fragmented, and continually evolving landscape of cloud solutions, we recommend only the best fit and value for your organization.

A LOOK AHEAD

Section 2.0
Differentiation: Introduces and diagrams the Architecture, and explains how it reflects Booz Allen’s unique approach. You’ll also read about our core design principles, extensive service offerings, and technology choices. Pages 10–15

Section 3.0
Depth: Takes the Architecture apart layer by layer with detailed visuals, design concepts, and recommended solutions from the cloud vendor landscape. The section ends with a look at how security is built into all levels. Pages 16–30

Section 4.0
Successes: Presents real-world examples from our extensive file of case studies. We present the solutions and challenges, describe and diagram the implementations, and explain the results. Pages 31–35
Booz Allen’s approach to cloud analytics is unmatched in the industry. Read about our unique principles and best practices.
A Layered Framework

The Booz Allen Cloud Analytics Reference Architecture incorporates a wide range of services to move from a technology infrastructure with chaotic, distributed data burdened by noise to large-scale data processing and analytics characterized by speed, precision, security, scalability, and cost efficiency.

However, Booz Allen’s approach is about much more than infrastructure. We start with your need to make better sense and better use of your mission data, and build from there.

Human Insights and Actions
Enabled by customizable interfaces and visualizations of the data

Analytics and Services
Your tools for analysis, modeling, testing, and simulations

Data Management
The single, secure repository for all of your valuable data

Infrastructure
The technology platform for storing and managing your data
Booz Allen
Cloud Analytics Service Offerings

**CLOUD STRATEGY AND ECONOMICS**
Delivery of strategy, technology, and economic analysis for evaluating and planning all of the business, technical, operational, and financial aspects of a cloud transition.

**CLOUD APPLICATION MIGRATION**
Expertise in the assessment, prioritization, architectural mapping, re-engineering, and optimization of workloads that have high value and are ready for migration to the cloud.

**ADVANCED CLOUD ANALYTICS**
Delivery of scalable analytics platforms allowing the processing of information at extreme scale; and eDiscovery: highvolume, full text indexing, and context-based search of information.

**INFRASTRUCTURE**
Design and implementation of IaaS offerings to provide global access to data storage, computing, and networking services on demand through self-service portals.

**CLOUD SECURITY**
Unified risk management approach to define cloud security requirements, controls, and a continuous monitoring framework to address data protection, identity, privacy, regulatory, and compliance risks.

**SOFTWARE AND PLATFORM**
Expertise in the secure implementation of SaaS and PaaS service delivery models, data migration, and integration with existing enterprise infrastructure and applications.

**VDI DEPLOYMENT AND INTEGRATION**
Delivery of flexible and dynamic virtual desktop infrastructure to simplify management, reduce licensing costs, and increase desktop security and data protection requirements.

**DATA CENTER MIGRATION AND OPTIMIZATION**
Identify critical factors, design, and execute the transformation of legacy IT systems to virtualized and cloud computing environments.
We’ll help you navigate the crowded, fragmented, and continually evolving vendor ecosystem to design a best-of-breed solution for your organization.
IN-SITU PROCESSING
The Architecture demands that “you send the question to the data,” because most big data processes are disk I/O-bound. In-situ processing means that most of the computation is done locally to the data, so that analytics run faster. This can enhance existing analytic capabilities and/or allow you to ask entirely new types of questions.

USE COMMODITY HARDWARE
Hardware should be expected to fail as the normal condition. The Architecture supports both scalability and fault tolerance to achieve optimal application load balancing.

SCHEMA ON READ
If you have all the source data indexed and queryable, plus the ability to create aggregations, then you can manage complex ontologies and demands in a very efficient manner.

THROW AWAY NOTHING
Near-linear scalable hardware and software systems allow much more data to be stored, which enable reprocessing of historical data with new algorithms and correlations that bring new insights.

DATA TAGGING
You can now afford to tag all of your data for sensitivity or other controls (such as geographic). This is the fastest, most reliable way to instrument change across your entire Data Lake.

ECONOMIES OF SCALE
What used to be called service-oriented architecture (SOA) means that you can define the value and cost of services in your enterprise, and plan your development actions, either to reduce the cost of low-value components or increase the scale of high-value components.

CHANGE DEVELOPMENT PROCESS
In order to develop a tight, iterative relationship with your end users, you can develop/research a new capability in hours (not months), and the process of discovery and integration with the rest of the enterprise begins much sooner, too.
Deeper Insights

The Cloud Analytics Reference Architecture enables staff at all levels to quickly gather and act on granular insights from all of your available data, regardless of its format or location. Below are some of the ways human insights and actions are enhanced by this new framework, which fosters greater collaboration and teamwork, and, ultimately, delivers the highest business value from your information and your computing infrastructure.

Human Insights and Actions

- Real-time alerting, situational awareness, and dissemination specific to their clearance level
- Investigate and provide feedback on reporting
- Interact and search using tailored tools

Analytics and Services

- Create and use many views into the same data
- Automatically find trends and outliers
- Evaluate analysis methods to determine and enhance best-of-breed tradecraft

Data Management

- No longer constrained by years-old schemas
- Catalog and index the data that is relevant today
- Free to create new views and reporting metrics
- Reference undiscovered trends in original data
- Apply advanced machine learning and statistical methods
- In-situ hypothesis testing

Infrastructure

- Reduce IT costs through commoditization and economies of scale
- Meet long-term scalability requirements

DEcisionMakers, INvestigators, InTERdictors, and ANALySTs

ANALySTs AND DATA SCIENTISTS

- blasted and diverse data management infrastructure

decisions, INVEStigators, INTERdictors, AND ANALySTs

- Real-time alerting, situational awareness, and dissemination specific to their clearance level
- Investigate and provide feedback on reporting
- Interact and search using tailored tools

ANALySTs AND DATA SCIENTISTS

- Create and use many views into the same data
- Automatically find trends and outliers
- Evaluate analysis methods to determine and enhance best-of-breed tradecraft

DEVELOPERS AND DATA SCIENTISTS

- No longer constrained by years-old schemas
- Catalog and index the data that is relevant today
- Free to create new views and reporting metrics
- Reference undiscovered trends in original data
- Apply advanced machine learning and statistical methods
- In-situ hypothesis testing

SYSTEM ADMINISTRATORS AND IT STAFF

- Reduce IT costs through commoditization and economies of scale
- Meet long-term scalability requirements
We diagram and describe each layer of the Cloud Analytics Reference Architecture, including our design principles and technology choices.
Booz Allen’s Cloud Analytics Reference Architecture provides a holistic approach to people, processes, and technology in four tightly integrated layers.

**Key Attributes**

By design, the Booz Allen Cloud Analytics Reference Architecture:

- Is reliable, allowing distributed storage and replication of bytes across networks and hardware that is assumed to fail at any time
- Allows for massive, world-scale storage that separates metadata from data
- Supports a write-once, sporadic append, read-many usage structure
- Stores records of various sizes, from a few bytes up to a few terabytes in size
- Allows compute cycles to be easily moved to the data store, instead of moving the data to a processor farm

**Human Insights and Actions**

Building on results and outputs from various analytical methods, multiple data visualizations can be created in your new cloud analytics solution. These are used to compose the interactive, real-time dashboard interfaces your decision-makers and analysts need to make sense of your data.

**Analytics and Services**

Both traditional and “Big Data” tools and software can operate on the information stored in your Data Lake, producing advanced specific analysis, modeling, testing, and simulations you need for decision making.

**Data Management**

Your Data Lake is a secure, distributed repository of a wide variety of data sources. Security, metadata, and indexing of Big Data are enabled by distributed key value systems (NoSQL), but the Architecture allows for traditional relational databases as well.

**Infrastructure**

This foundational layer allows for quick, streamlined, low-risk deployment of the cloud implementation. The plug-and-play, vendor-neutral framework is unique to Booz Allen.

---

*A FRAMEWORK FOR SECURITY*

Page 30 details our security processes
Human Insights and Actions

ARCHITECTURE MODEL

Human Insights and Actions

Interfaces

<table>
<thead>
<tr>
<th>Monitoring</th>
<th>Alert</th>
<th>CCH</th>
<th>Exploratory</th>
</tr>
</thead>
</table>

Visualizations

<table>
<thead>
<tr>
<th>Histogram</th>
<th>Pie Chart</th>
<th>Bar Chart</th>
<th>Geospatial</th>
</tr>
</thead>
<tbody>
<tr>
<td>Networks and Graphs</td>
<td>Control Chart</td>
<td>Scatter Plot</td>
<td>Line Chart</td>
</tr>
<tr>
<td>Time Series</td>
<td>Tree Maps</td>
<td>Word Cloud</td>
<td>Parallel Coordinates</td>
</tr>
<tr>
<td>ROC</td>
<td>Survival Chart</td>
<td>Box and Whisker</td>
<td></td>
</tr>
</tbody>
</table>

MONITORING

EXPLORATORY

GEOSPATIAL

LINE CHART
In analytics solutions built on the Architecture, the data that’s available and the desired results drive the interfaces—not the other way around. When user communities and stakeholders aren’t restricted by their tools, they can perform complex visualizations to identify patterns they previously couldn’t see.

That freedom defines the guiding principles behind this first layer of the Architecture:

- Design and build the framework so that the desired data and analytic results define the visualization
- Reuse results and outputs of analytics across different visualizations
- Decouple the underlying analytics and data access from the visualizations and interfaces so that it’s possible to build customized, interactive dashboard interfaces composed of dynamically linked visualizations

### TECHNOLOGY EXAMPLES

<table>
<thead>
<tr>
<th>Technology</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HTML5, JavaScript, OWF, Synapse</td>
<td>Lightweight, custom web-based applications and dashboards tailored to specific user communities or stakeholders for data exploration, event alerting, and monitoring, as well as continuous quality improvement</td>
</tr>
<tr>
<td>Commercial products (Splunk, Pentaho, Datameer Business Infographics, etc.)</td>
<td>Out-of-the-box, easy-to-build dashboards for historical trending and real-time monitoring to analyze user transactions, customer behavior, network patterns, security threats, and fraudulent activity</td>
</tr>
<tr>
<td>Adobe Flex and Adobe Flash</td>
<td>Despite the rise of HTML5, Adobe Flex and Flash applications still remain strong candidates for quickly building and deploying rich user interfaces</td>
</tr>
</tbody>
</table>
Analytics and Services

ARCHITECTURE MODEL

Human Insights and Actions

Data Mining, Text Mining, Machine Learning, Statistics

<table>
<thead>
<tr>
<th>Time Series</th>
<th>Summarization</th>
<th>Genetic Algorithms</th>
<th>Social Network Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anomaly Detection</td>
<td>Collaborative Filtering</td>
<td>Entity Extraction</td>
<td>Line Chart</td>
</tr>
<tr>
<td>Classification</td>
<td>Decision Trees/Rules</td>
<td>Topic Modeling</td>
<td>Regression</td>
</tr>
<tr>
<td>Simulation</td>
<td>Clustering</td>
<td>Association Rules</td>
<td></td>
</tr>
</tbody>
</table>

Data and Analytical Tools and Software

| R, SAS, Matlab, Mathematica | MapReduce, Hive, Pig, Hama | ETL |

TIME SERIES

SOCIAL NETWORK ANALYSIS

R, SAS, MATLAB, MATHEMATICA

MAPREDUCE, HIVE, PIG, HAMA
Frequently where data is concerned, the whole is greater than the sum of its parts. In the most strategic business decisions, the ability to combine multiple types of analyses creates a holistic picture that can lead to much more valuable insight. With the Cloud Analytics Reference Architecture, you can implement different types of analytical methods.

This integrated approach is an anchor for the guiding principles behind our Analytics and Services layer:

- Allow both traditional and Big Data analysis tools and software to operate on a centralized repository of data (the DataLake)
- Integrate results and outputs of analyses and visualize them on dashboards for decision making
- Decouple tools from the various types of analyses to make the system more extensible and adaptable
- Include a service-oriented architecture layer to reuse results and outputs in many different ways relevant to different stakeholders and decisionmakers
- Incorporate Certified Catastrophe Risk Analysis (CCRA) to allow a variety of data analysis tools and software to be integrated and used; it also enables results and outputs of analyses to be visualized and used across multiple interfaces

### TECHNOLOGY EXAMPLES

<table>
<thead>
<tr>
<th>Data Mining</th>
<th>Data mining is used to discover patterns in large datasets and draws from multiple fields including artificial intelligence, machine learning, statistics, and database systems.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Machine Learning</td>
<td>Machine learning is used to learn classifiers and prediction models in the absence of an expert and employs many algorithms in the areas of decision trees, association learning, artificial neural networks, inductive logic programming, support vector machines, clustering, Bayesian networks, genetic algorithms, reinforcement learning, and representation learning.</td>
</tr>
<tr>
<td>Natural Language Processing (NLP)</td>
<td>NLP is used to process unstructured and semi-structured documents for the purposes of information retrieval, sentiment analysis, statistical machine translation, and classification.</td>
</tr>
<tr>
<td>Network Analysis</td>
<td>Network analysis using graph theory and social network analysis are used to understand association and relationships between entities of interests.</td>
</tr>
<tr>
<td>Statistical Analysis</td>
<td>Traditional statistical methods using univariate and multivariate analysis on relatively small datasets are employed to make inferences, test hypotheses, and summarize data.</td>
</tr>
</tbody>
</table>
Analytics and Services (continued)

DISCOVERING YOUR DATA

Before working with Booz Allen, most clients faced a fundamental challenge with data discovery. They didn’t know what data was actually available or how to sort through all of it to identify the most important business problems or trends it could reveal.

TECHNICAL FRAMEWORK

Discovery is intimately related to search and analysis. All three feed into insight in a nonlinear fashion. A search-discovery-analytics process that solves business problems without consuming disproportionate resources meets these user needs:

▶ Real-time, ad hoc access to content
▶ Aggressive prioritization based on importance to the user and the business
▶ Data-driven decision making, which relies on the ability to try different approaches and ideas in order to discover previously unimagined insights
▶ Feedback/learning from the past intelligently applied to today’s data

HOW BOOZ ALLEN SIMPLIFIES DISCOVERY

Other solutions require analysts to break down data into numerous subsets and samples before it can be digested. This expensive, time-consuming process is one of the major roadblocks to turning data into true business intelligence.

Even though the Booz Allen Cloud Analytics Reference Architecture supports the most advanced analysis, it can also allow your staff to sift through all of your data on a basic level. Without tedious or sophisticated sampling and complex tools, they can discover what’s useful and what’s not useful for a specific business problem.

How does the Architecture support fast, efficient, and scalable search on entire datasets, not just samples?

▶ Bulk and soft real-time indexing enable the solution to handle billions of records with subsecond search and faceting
▶ Large-scale, cost-effective storage and processing capabilities accommodate “whole data” consumption and analysis; in-memory caching of critical data ensures applications meet performance requirements
▶ NLP and machine learning tools can scale to enhance discovery and analysis on very large datasets
How the Data Science Lifecycle Distills Insights

The data science lifecycle consists of three basic steps:

**Step 1**
First, data is sampled using a cloud analytics platform. This step may involve a sophisticated analytic that runs in the cloud, such as one that crawls a social network to find people with certain types of relationships with an individual or organization. This sampling can be done using either high-level query languages that are specially made for scalable cloud analytics or low-level developer interfaces.

**Step 2**
Next, a data scientist models the data sample in order to understand it better. This is usually done using a statistical modeling environment on the data scientist’s workstation.

**Step 3**
Finally, once a trend is established using the model, the data scientist works with analysts and domain experts to explain the trend and yield insights.

This cycle is repeated until the data science team reaches actionable insights and intelligence that can be presented to senior leadership for decision making purposes. Information may be delivered in a visualization, dashboard, or written report.
# Data Management

## ARCHITECTURE MODEL

### Data Management

<table>
<thead>
<tr>
<th>Data Lake</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Security Tagging</td>
<td>Metadata</td>
<td>Indexing</td>
<td>NoSQL: Key-Value</td>
</tr>
<tr>
<td>HDFS</td>
<td>RDBMS</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Sources</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Streaming</td>
<td>Batch</td>
<td>Text</td>
<td>Structured</td>
</tr>
<tr>
<td>Unstructured</td>
<td>Binary</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A central feature of the Architecture, the Data Lake delivers on the promise of cloud analytics to offer previously hidden insights and drive better decisions. It’s a secure repository for data of all types and origins. Instead of precategorizing data, which restricts its usability from the moment it enters your organization, the Architecture combines unstructured, structured, and streaming data types and makes them available for many different forms of analysis.

The following principles demonstrate how the Architecture enables your organization to use this repository of enterprise data to the best advantage:

- Provide inherent replication of the data through a distributed file system
- Use distributed key value (NoSQL) data storage to enable security and metadata tagging at the data level as well as indexing for specialized retrieval
- Relax schema constraints and provide the flexibility to adapt to changing data sources and types with the schema-on-read approach of distributed key value data storage
- Store the Data Lake on commodity hardware and scale linearly in performance and storage
- Don’t presummarize or precategorize data
- Enable rapid ingest of data, aggressive indexing, and dynamic question-focused datasets through scale

**TECHNOLOGY EXAMPLES**

<table>
<thead>
<tr>
<th>Technology</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hadoop Distributed File System (HDFS)</td>
<td>The primary open-source, distributed storage system creates multiple replicas of data blocks and distributes them on compute nodes throughout a cluster to enable reliable, rapid computations.</td>
</tr>
<tr>
<td>Accumulo</td>
<td>NoSQL store based on Google’s BigTable design features cell-level security access labels and a server-side programming mechanism that can modify key/value pairs at various points in the data management process.</td>
</tr>
<tr>
<td>Hbase, Cassandra, MongoDB</td>
<td>Open-source NoSQL databases focused on a combination of consistency, availability, and partition tolerance.</td>
</tr>
<tr>
<td>Neo4j</td>
<td>NoSQL scalable graph database storing data in nodes and the relationships of a graph.</td>
</tr>
</tbody>
</table>
Data Management (continued)

Booz Allen works with organizations in corporate and government sectors that have an urgent need to make sense of volumes of data from diverse sources, including those that had been inaccessible or extremely difficult to utilize, such as streams from social networks. Now analysts and decisionmakers can form new connections between all of this data to uncover previously hidden trends and relationships.
Booz Allen’s strategy and technology consultants are highly regarded subject matter experts. Through groundbreaking conference keynotes, whiteboard talks, and papers, they help educate and shape the analytics industry.

We invite you and your team to take advantage of the educational resources listed below to gain strategic insights about the use of analytics, explore technical topics in depth, and stay on top of the latest trends.

**Presentations**

**Yahoo! Hadoop Summit: Biometric Databases and Hadoop**
Invented and demonstrated methods for dense data correlation (e.g., imagery and biometrics) within a Hadoop distributed computing platform using new machine learning parallel methods.

**Yahoo! Hadoop Summit: Culvert—A Robust Framework for Secondary Indexing of Structured and Unstructured Data**
Demonstration of Booz Allen’s secondary indexing solutions and design patterns, which support online index updates as well as a variation of the HIVE query language over Accumulo and other BigTable-like databases to allow indexing one or more columns in a table.

**Slidecast: Hadoop World—Protein Alignment**
Demonstration of advanced analytics in using protein alignment sequences to identify disease markers using Hadoop, HBase, Accumulo, and novel machine learning concepts.

**Slidecast: Innovative Cyber Defense with Cloud Analytics**
Presentation on improving intelligence analysis through a hybrid cloud approach to analytics, with descriptions and diagrams from Booz Allen client solutions.

**Slidecast: Integrating Tahoe with Hadoop’s MapReduce**
Invented and demonstrated method to use least-authority encrypted file system as plugin to HDFS within Hadoop cluster.

**Papers**

**Massive Data Analytics in the Cloud**
Overview of the business impact of cloud computing, and how data clouds are shaping new advances in intelligence analysis.

**Videos**

**Cloud Whiteboard Playlist**
Short instructional videos on a range of topics from introductory talks for executives to tutorials for data analysts. Check back frequently for new material.

**Cloud Analytics for Executive Leadership**
Booz Allen Principal Josh Sullivan discusses how analysis of data can be used as a tool to provide insight to executives.

**Informed Decision Making: Sampling Techniques for Cloud Data**
Booz Allen Data Scientist Ed Kohlwey explains how sampling large amounts of data can be useful for program managers to make informed decisions.

**Developer Perspectives: The FuzzyTable Database**
Booz Allen Data Scientist Drew Farris explains how to use the FuzzyTable biometrics database.

**Workshop**

**O’Reilly Strata Conference: Beyond MapReduce—Getting Creative with Parallel Processing**
Technical discussion of MapReduce as an excellent environment for some parallel computing tasks and the many ways to use a cluster beyond MapReduce.
# Infrastructure

## ARCHITECTURE MODEL

### Core Services Layer

<table>
<thead>
<tr>
<th>Application/Data Services</th>
<th>Interfaces</th>
<th>API</th>
</tr>
</thead>
<tbody>
<tr>
<td>Platform Services (PaaS)</td>
<td>Software Platform</td>
<td>APIs and Tools</td>
</tr>
<tr>
<td>Infrastructure Services (IaaS)</td>
<td>Cluster Controller</td>
<td>OS Inventory</td>
</tr>
</tbody>
</table>

### Resources Abstraction Layer

<table>
<thead>
<tr>
<th>Virtual Machine</th>
<th>Virtual Network</th>
<th>Virtual Storage</th>
<th>Abstraction Manager</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guest OS</td>
<td>Virtual Network Services</td>
<td>File System</td>
<td>Discovery</td>
</tr>
<tr>
<td>VM Tools</td>
<td>API</td>
<td>Interfaces</td>
<td>Provisioning</td>
</tr>
<tr>
<td>Hypervisor</td>
<td>Kemal</td>
<td>Drivers</td>
<td>Orchestration</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bus</td>
<td>API</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Security</td>
</tr>
</tbody>
</table>

### Physical Resource Layer

<table>
<thead>
<tr>
<th>Computing Resource Pool</th>
<th>CPU</th>
<th>Memory</th>
<th>Networking</th>
<th>Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hosting Platform</td>
<td>Facility</td>
<td>Racks</td>
<td>Servers</td>
<td>Power</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resource Manager</th>
<th>Acceleration</th>
<th>Security</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Infrastructure is the foundation for any cloud implementation. What makes the Booz Allen Cloud Analytics Reference Architecture unique is its plug-and-play, vendor-neutral framework. This framework not only allows a greater range of choices in selecting resources and building services, it also allows for a faster, more streamlined, more secure, and lower risk deployment.

The following principles guide the infrastructure layer of the Architecture:

▶ Make it easy to transform physical resources from legacy IT systems to secure, virtualized data centers and trusted cloud computing environments
▶ Implement core services to provide the mechanisms to realize on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service
▶ Employ virtualization to increase utilization of existing assets and resources, and improve operational effectiveness
▶ Engineer in-depth security to provide controls and continuous monitoring in order to fully address data protection, identity, privacy, regulatory, and compliance risks

TECHNOLOGY EXAMPLES

Amazon Web Services, Microsoft Azure, Puppet, VMware, vSphere

Cloud tool chain for provisioning, configuration, orchestration, and monitoring of virtual environment. These tools provide the building blocks for IaaS, PaaS, and foundation for SaaS. Run multiple operating systems and virtual network platforms on the same hardware—sharing computing, storage, and networking resources.

Security through VMware, McAfee, Symantec, Cisco, TripWire, EnCase

Protect assets—physical, logical, and virtual—while automating governance and compliance.
## Reference Architecture

### Security Framework

The Architecture is designed to protect your data at rest and in flight, with security controls embedded in each layer. This is obviously more than just a technology challenge. We understand the need to embed new processes and training regimens so your staff handles sensitive data correctly. We also advise you on how to secure your facilities and ensure that all off-premise facilities have the right controls in place as well.

<table>
<thead>
<tr>
<th>Business</th>
<th>Geography</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assets to Be Protected</td>
<td>Distributed Sites</td>
</tr>
<tr>
<td>Threats and Processes that Require Security</td>
<td>Time Dependencies</td>
</tr>
<tr>
<td>Organizational Security</td>
<td>Governance</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Conceptual</th>
<th>Technical and Management Security Strategies</th>
<th>Roles and Responsibilities</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Attributes</td>
<td>Trust Relationships</td>
<td>Time Dependencies</td>
</tr>
<tr>
<td>Security Requirements to Support the Business</td>
<td>Security Domains, Boundaries, and Associations</td>
<td>When Is Protection Relevant?</td>
</tr>
<tr>
<td>Control and Enablement Objectives Resulting from Risk Assessment</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Logical</th>
<th>Interrelationships</th>
<th>Security Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Information to Be Secured</td>
<td>Attributes</td>
<td>Authentication Confidentiality and Integrity Protection</td>
</tr>
<tr>
<td>Security and Risk</td>
<td>Management Policy</td>
<td>Strategic Partnerships</td>
</tr>
<tr>
<td>Security Entities</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Physical</th>
<th>Security Mechanisms</th>
<th>Human Interface</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security-Related Data Structures</td>
<td>Encryption</td>
<td>Screen Formats</td>
</tr>
<tr>
<td>Tables</td>
<td>Access Control</td>
<td>User Interaction</td>
</tr>
<tr>
<td>Messages</td>
<td>Digital Signatures</td>
<td>Access Control</td>
</tr>
<tr>
<td>Pointers</td>
<td>Security Infrastructure</td>
<td>Systems</td>
</tr>
<tr>
<td>Certificates</td>
<td>Physical Layout of Hardware, Software, and Communication Lines</td>
<td>Time Dependencies</td>
</tr>
<tr>
<td>Signatures</td>
<td></td>
<td>Sequence of Processes and Sessions</td>
</tr>
<tr>
<td>Security Rules</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conditions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practices</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Procedures</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Component</th>
<th>Personnel Management Tools</th>
<th>Time Dependencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Security IT Products</td>
<td>Identities</td>
<td>Time Schedules</td>
</tr>
<tr>
<td>Risk Management</td>
<td>Roles</td>
<td>Clocks</td>
</tr>
<tr>
<td>Tools for Monitoring and Reporting</td>
<td>Functions</td>
<td>Timers and Interrupts</td>
</tr>
<tr>
<td>Security Process</td>
<td>Access Controls</td>
<td></td>
</tr>
<tr>
<td>Tools, Standards, and Protocols</td>
<td>Lists</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Service Management</th>
<th>Management of Security Operations</th>
<th>Management of Environment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service Delivery Management</td>
<td>Admin</td>
<td>Buildings</td>
</tr>
<tr>
<td>Assurance of Operational Continuity</td>
<td>Backups</td>
<td>Sites</td>
</tr>
<tr>
<td>Operational Risk Management</td>
<td>Monitoring</td>
<td>Platforms and Networks</td>
</tr>
<tr>
<td>Risk Assessment</td>
<td>Emergency Response</td>
<td>Management Schedule</td>
</tr>
<tr>
<td>Monitoring and Reporting</td>
<td>Personnel Management</td>
<td>Security-Related</td>
</tr>
<tr>
<td></td>
<td>Account Provisioning</td>
<td>Calendar and Timetable</td>
</tr>
<tr>
<td></td>
<td>User Support</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Management</td>
<td></td>
</tr>
</tbody>
</table>

Reference Architecture

A FRAMEWORK FOR SECURITY

The Architecture is designed to protect your data at rest and in flight, with security controls embedded in each layer. This is obviously more than just a technology challenge. We understand the need to embed new processes and training regimens so your staff handles sensitive data correctly. We also advise you on how to secure your facilities and ensure that all off-premise facilities have the right controls in place as well.

---

**A FRAMEWORK FOR SECURITY**

The Architecture is designed to protect your data at rest and in flight, with security controls embedded in each layer. This is obviously more than just a technology challenge. We understand the need to embed new processes and training regimens so your staff handles sensitive data correctly. We also advise you on how to secure your facilities and ensure that all off-premise facilities have the right controls in place as well.
These case studies show how Booz Allen uses superior technology and analytics expertise to solve complex problems for clients in a wide range of corporate and government sectors.
Improving Intelligence Analysis

Mission
To fulfill their mission, this organization requires data correlation, quick access to analytic results, ad-hoc queries, advanced scalable analytics, and real-time alerting. To provide their analysts with a continuous pipeline of prioritized, actionable information, they needed a secure, scalable, automated solution that would more quickly and precisely sift through large (and growing) volumes of complex data characterized by a variety of formats and noise. In addition, they needed to leverage their existing analytics infrastructure in the new platform.

Solutions
Booz Allen worked closely with the client to adopt a data cloud implementation by augmenting the legacy relational databases with cloud computing and analytics. The design focused on keeping transactional-based queries in the current relational databases, while doing the “heavy lifting” in the cloud and outputting the interesting, processed, or desired analytic results into relational data stores for quick transactional access. With many existing systems and applications dependent on the legacy relational database for transactional queries of data, Booz Allen pulled together excess servers from the client’s infrastructure to build a hybrid cloud solution. Also, as the client’s needs change to adapt to the mission, the solution is scalable and flexible to support future innovation and evolution without reengineering.

Booz Allen worked closely with the client to adopt a data cloud implementation by augmenting the legacy relational databases with cloud computing and analytics. The design focused on keeping transactional-based queries in the current relational databases, while doing the “heavy lifting” in the cloud and outputting the interesting, processed, or desired analytic results into relational data stores for quick transactional access.

Interfaces and Visualizations
Dashboards, web applications, client applications, and rich clients interfaced and integrated with advanced analytics infrastructure and legacy relational databases through a SOA business logic layer.

Analytics and Services
The solution called for predictive analytics to forecast potential events from existing data and anomaly detection to extract potentially significant information and patterns. The solution leveraged the core principle of cloud analytics that enables automated analysis techniques, precomputation, and aggressive indexing.

Data Management
The data sources had multiple formats, were large in size, and distressed with noise. The solution created deep insight through fusion of different data types at scale. The solution enabled the ability to follow the lineage or pedigree of the data, allowing the client to map cost in relation to the value of the data or how well it is being used.

Infrastructure
The solution used Accumulo (distributed key value systems/NoSQL database) for content normalization and indexing, MapReduce as the precomputation engine, and HDFS for scalable ingest and storage.

Impact
Rather than simply focus on gaining IT efficiencies by using cloud technology for infrastructure, Booz Allen focused on applying cloud analytics and in-depth understanding of the organization’s operational and mission needs to extract more value faster from massive datasets. The new cloud solution provided immediate and striking improvements across the increasing volume of structured and unstructured data using aggressive indexing techniques, on-demand analytics, and precomputed results for common analytics.

The final solution combined sophistication with scalability, moving the organization from a situation in which analysts stitched together sparse bits of data to a platform for distilling real-time, actionable information from the full aggregation of data.
Planning and Responding to Disaster

Mission
This organization, which is responsible for disaster planning and response, found that social media could provide timely situational awareness for biological (and other disaster) events. They wanted a solution to better characterize and forecast emerging disaster events using social media data as it streams in real time. With such a solution in place, the organization could increase overall preparedness by leveraging event characterization to accurately predict the impact and improve the response.

In order to reach their goal, the organization needed higher levels of confidence in the social media data on which they would base their decisions. The specific challenges the new solution had to overcome included data ingestion and normalization, social media vocabulary, social media characterization, information extraction, and geographical isolation of events.

Solutions
Booz Allen developed a framework to capture, normalize, and transform open-source media used to characterize and forecast disaster events, in real time. The framework incorporated computational and analytical approaches to turn the noise from social media into valuable information using algorithms such as term frequency-inverse document frequency (TF-IDF), natural language processing (NLP), and predictive modeling to characterize and forecast the numbers of sick, dead, and hospitalized, as well as to extract symptoms, geography, and demographics for specific illness events.

The solution framework was implemented in the cloud, taking advantage of the flexible computational power and storage. The new cloud infrastructure allowed Booz Allen’s data capturing and visualization tool, Splunk, to mine through and analyze vast amounts of data in real time, while outputting characterization and forecasting metrics of captured events.

Interfaces and Visualizations
The solution included dashboards that characterized events captured in social media. The visual analyses include event extraction counts, time series counts, forecasting counts, a symptom tag cloud, and geographical isolation.

Analytics and Services
TF-IDF and NLP algorithms were used to classify and extract relevant information from the data. Booz Allen developed predictive models for forecasting event frequency and counts. The algorithms were written in Python and incorporated into Splunk located on Amazon Web Services (AWS).

Data Management
The solution framework captured live, streaming open-source media such as Twitter and RSS feeds. Data was captured in Splunk and stored on AWS.

Impact
The new Booz Allen solution, which builds upon current best practices in cyber terrorism, enables near real-time situational awareness through a standalone surveillance system that captures, transforms, and analyzes massive volumes of social media data. By leveraging social media data and analytics for more timely and accurate disaster characterization, the organization is able to more effectively plan and respond.
Detecting Fraud and Abuse

Mission
Medicare and Medicaid pay out approximately $750 billion each year to more than 1.5 million doctors, hospitals, and medical suppliers. By many estimates, about $65 billion a year is lost to fraud. This organization needed to be able to detect fraud in claim data streams and stop processing immediately; they also wanted to assign a fraud risk score to providers and patient data in order to prioritize their investigations. They were challenged by multiple disparate sources of data, including valuable historic data archived in currently inaccessible formats.

In addition, fraud and abuse techniques are evolving rapidly, as are policies and technologies, so the final solution could not lock them into specific tools, data sources, or approaches to detection. Lastly, the solution had to allow them to operate in compliance with regulatory requirements and laws governing the use of personally identifiable information.

Solution
Booz Allen used a variety of analytical techniques and detection methods to support the creation and maintenance of tools that allow organizations to stay ahead of criminals. The solution for this client integrates and combines the best technologies and analytics available to enable the analysis of multiple data sources. Booz Allen built systems for routine detection that are designed to accept new data sources and techniques for detection. For example, Booz Allen helped build a risk-scoring algorithm that drew information from multiple federal and civil data sources. The risk-scoring system is flexible enough to allow analysts to build new rules quickly, and the cloud architecture can then accurately rescore the entire population.

Interfaces and Visualizations
Users are given an interface to monitor overall provider risk. They can drill down into data on each provider to get more statistical information and visualizations to gain insight into specific risk factors and to compose forecasts.

Analytics and Services
Geotagging, risk scoring, and predictive modeling analysis are applied to the data. Specific predictive analyses include neural nets, clustering, and regression. A rule-based system is also used to detect many of the known kinds of fraud.

Impact
For the first time, doctors and others who want to bill Medicare are being assessed based on their risk to commit fraud. Those who are deemed likely to commit fraud or have a record of investigations are rooted out. In addition, payers can better prioritize and target investigations to prevent improper payment or to recover funds.
Predicting and Detecting Disease

Mission
This organization is charged with evaluating and measuring the efficacy of hospital compliance with SSC guidelines for addressing Severe Sepsis and Septic Shock (S4). They needed to develop a new solution for compliance analysis and early detection analysis in order to lower mortality rates and overall health costs related to S4.

The final solution needed to allow them to mine Electronic Health Records (EHR) for clinical indicators that could lead to early detection of S4 and predict the development of S4 from sepsis. They also wanted to enable hospitals to harness the value of patient information to diagnose more quickly, and use this data to decrease the time between official diagnosis and implementation of the standard of care.

Solution
Booz Allen’s team led a cross-company project, Sepsis Intervention Outcomes Research (SIOR), that tapped analytical, clinical, economic, and informatics expertise. SIOR analyzed medical workers’ compliance with international standards of care for S4, and compared that compliance with patient outcomes. Booz Allen’s advanced analytics experts helped develop an Event-Centric Ontology (ECO) that incorporated NLP of medical personnel notes. ECO provided a formalized vocabulary and framework for evaluating EHRs that expedited real-time discovery and harnessing of structured and unstructured data. Booz Allen also developed a predictive model based on vital measurements at critical times to produce a risk score for developing S4 from sepsis. In addition, Mahalanobis distance plots from baseline showed that signals are present before POD, which allows for earlier detection of at-risk patients.

Analytics and Services
The solution used logistic regression, NLP correlation, and time-series analyses.

Data Management
Booz Allen obtained over 27,000 individual patient EHRs for analysis containing both structured and unstructured data spanning four hospitals and a period of 2 years.

Impact
Compliance analysis suggests a strong correlation between compliance to SSC guidelines and decreased mortality. Early detection analysis indicates there may be a set of clinical indicators that could be used to identify patients at risk for developing S4, allowing their care to be prioritized. Booz Allen developed an analytically expedient framework that allows for more efficient computation and discovery of underlying relationships, which can allow hospitals to expedite diagnosis and treatment and save more lives.
ABOUT BOOZ ALLEN

Booz Allen Hamilton has been at the forefront of strategy and technology consulting for nearly a century. Today, the firm provides services primarily to the US government in defense, intelligence, and civil markets, and to major corporations, institutions, and not-for-profit organizations. Booz Allen offers clients deep functional knowledge spanning strategy and organization, engineering and operations, technology, and analytics—which it combines with specialized expertise in clients’ mission and domain areas to help solve their toughest problems.

Booz Allen is headquartered in McLean, Virginia, employs approximately 25,000 people, and had revenue of $5.86 billion for the 12 months ended March 31, 2012. To learn more, visit www.boozallen.com. (NYSE: BAH)