November 8, 2018

TRENDS

How to Blend Self-Service and Solid Governance for the Hybrid Data Lake

Balance Accessibility and Control Across Cloud and On-Premises Big Data Deployments

Doug Henschen

Vice President and Principal Analyst

Copy Editor: Jim Donahue

Layout Editor: Aubrey Coggins

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Leading organizations pursuing digital transformation are turning to big data and cloud deployments to drive agile development and innovation. Data lakes, Internet of Things initiatives, artificial intelligence and machine learning experiments as well as self-service analytics programs are all moving into the cloud. Yet even the most aggressive companies that are "all in" on the cloud often choose to retain certain data and related assets on-premises because of privacy or other regulatory requirements. Trust is a core concern in any data initiative, yet governance and assurance of compliance have never been more challenging now that organizations have data and assets spread across the cloud and on-premises data centers.

This report explains how organizations can deliver fast and business-user-friendly self-service access to information while also ensuring proper data governance. It discusses data catalogs, which are proliferating because they ease the data-access problem for business users. But as this report explains, catalogs must include or tightly integrate with data-lineage tracking, data glossaries and broader data-governance capabilities. What's more, organizations must address the people and process aspects of data governance to ensure user adoption and trustworthy information.

Business Themes


Data to Decisions


Technology Optimization

DIGITAL TRANSFORMATION RUNS ON BIG DATA AND CLOUD

Here's the paradox: Data has become the lifeblood of organizations, powering new business models, driving customer experiences and accelerating business decision-making. Yet the task of ensuring that data is accessible, secure and well-governed has never been more challenging. Here's why:

- Data-generating applications have multiplied and spread to the cloud and mobile devices and edge sensors.
- Enterprises now routinely manage data lakes measured in tens to hundreds of terabytes if not petabytes, and unstructured and variable data is almost always in the mix.
- Infrastructure, storage and processing options have multiplied, as have on-premises, virtualized and cloud-based deployments.
- Data-privacy laws are sprouting up around the globe while regulatory requirements are more demanding than ever.

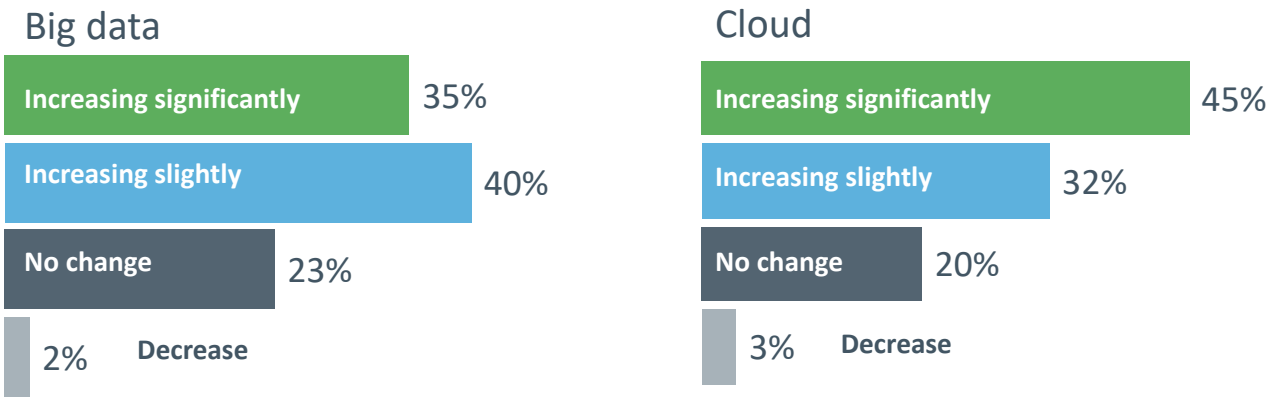
Against this backdrop, organizations must improve customer experiences and drive growth or risk being displaced by more agile and innovative competitors. Nearly 70 percent of organizations have digital transformation strategies in place, according to an October 2017 study by Constellation Research.[1] C-level executives tell us the top technology enablers of transformation are analytics and big data, cited by 75 percent of respondents, followed by public cloud, cited by 67 percent of respondents.[2]

The big data and cloud trends have each been growing for nearly a decade and show no sign of abating. In fact, many data lakes are destined to get exponentially larger, thanks to a flood of new information generated by Internet of Things (IoT) initiatives. Machine leaning (ML) and artificial intelligence (AI) programs, meanwhile, will depend on cloud scale and agile computing capacity. It's no wonder that a clear majority of respondents say they'll increase their investments on these two fronts over the coming year (see Figure 1).

Figure 1. Big Data and Cloud Are Top Tech Enablers of Digital Transformation

What is your company's plan for investment in big data and cloud in the next year?

Investments in big data and cloud technology are growing to support digital transformation.



Source: Constellation Research[3]

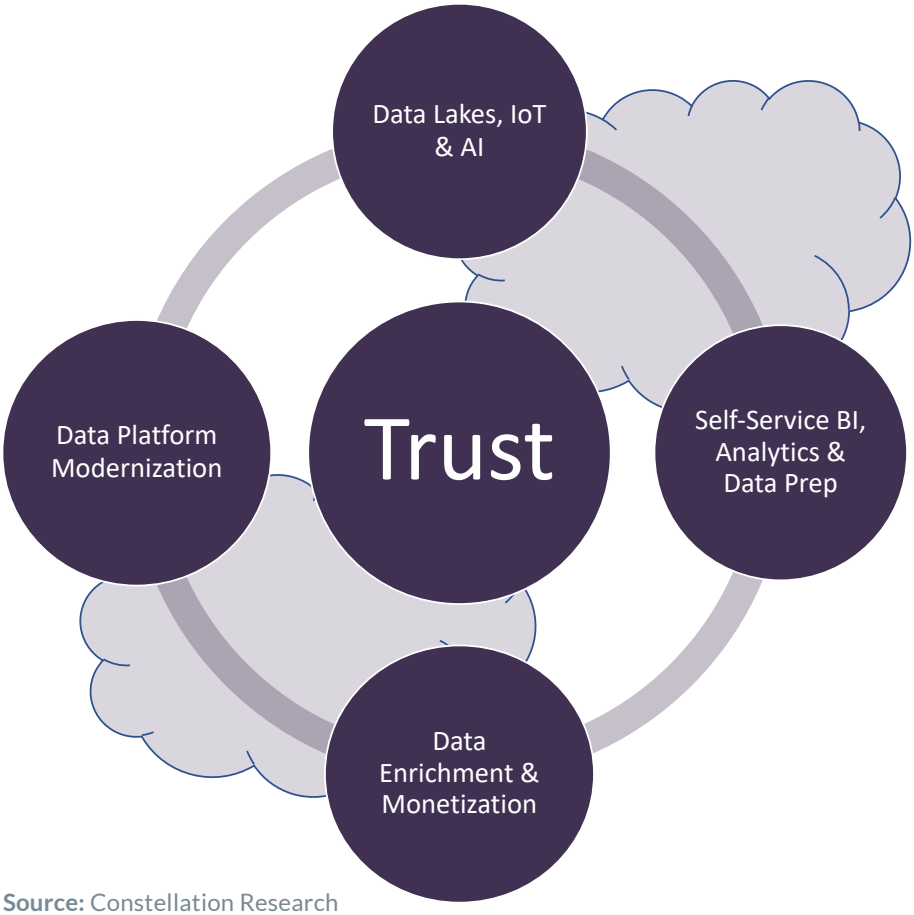
TRUST IS THE CORE REQUIREMENT FOR ANY DATA INITIATIVE

Data lakes aren't the only data initiatives being both fueled and complicated (from a governance perspective) by cloud migrations and deployments (see Figure 2). The drive toward self-service business intelligence and analytics, for example, is expanding to include self-service data preparation. Ensuring easy access to data is a prerequisite for any form of self-service, yet access complicates trust challenges ranging from security and access control to data quality, data certification and tracking of data lineage.

The push to innovate and drive new business models is spawning data-enrichment and monetization programs. Location, weather and psychographic data, for example, is being used to bring deeper insight and personalization to customer services. These moves introduce yet more data and, often, third-party data that complicates data management and governance. Ditto cloud services and architecture, which support the delivery of data and decision services but also complicate governance challenges such as tracking data lineage and enforcing data-usage policies.

All data initiatives draw on data platforms, and here's where modernization initiatives present a moving target. As noted in Figure 1, that motion is headed in one direction: the cloud. Cloud-based database

Figure 2. Governance and Trust Are Tougher Challenges in a Hybrid World



Source: Constellation Research

services, Hadoop and Spark services and, more recently, cloud-based streaming services and object stores are shouldering a growing share of the data-platform burden. But how best to ensure data visibility and appropriate access as well as solid governance, including tracking of data lineage and enforcement of data policies across hybrid, cloud and on-premises data landscapes?

DATA CATALOGING AND GOVERNANCE EVOLVE TO SUPPORT AGILITY AND TRUST

The good news is that technologies have emerged that deliver visibility into data and related assets—whether on-premises or in the cloud—while also ensuring comprehensive data governance. To provide agile data access, data catalogs have quickly gained traction as the way to make data more

visible to business users. The question is whether the catalogs provide a comprehensive view of data across hybrid deployments and to what degree they also address data governance needs, such as standardization of metadata, tracking of lineage, stewardship and enforcement of data policies.

Catalogs that can span all data sources and deployment choices are a great start to providing comprehensive data access and to addressing trust, but many catalogs do not support all sources or deployment choices. In other cases, catalogs span deployment options but do not address the data-governance functions shown below the dotted line in Figure 3. Here's where evolved, integrated options come in that combine the automation and ML capabilities of catalogs while also addressing data-governance needs. The most effective catalogs not only show users what data is available but also expose governance-related nuances—such as quality, lineage and policy constraints—to provide complete context and a better understanding of the best and most appropriate uses of data and related assets.

Let's take a closer look at what to look for when adopting data catalogs and integrated data-governance capabilities.

Data Catalogs Ease Access, Collaboration

It's no coincidence that data catalogs have gained popularity along with the rise of data lakes. Many early data lakes more closely resembled swamps, in that data sets were dumped into the lake with little thought toward organization or minimal creation of zones for different types of data or stages of data preparation. Catalogs emerged to help organizations create an inventory of available data assets within a lake. As a practical matter, most organizations ended up deploying multiple data lakes instead of a single data lake for the sake of expediency, security and simplicity. Thus, leading catalogs were quickly adapted to span multiple data lakes as well as conventional relational sources and supporting systems, such as ETL and Spark environments.

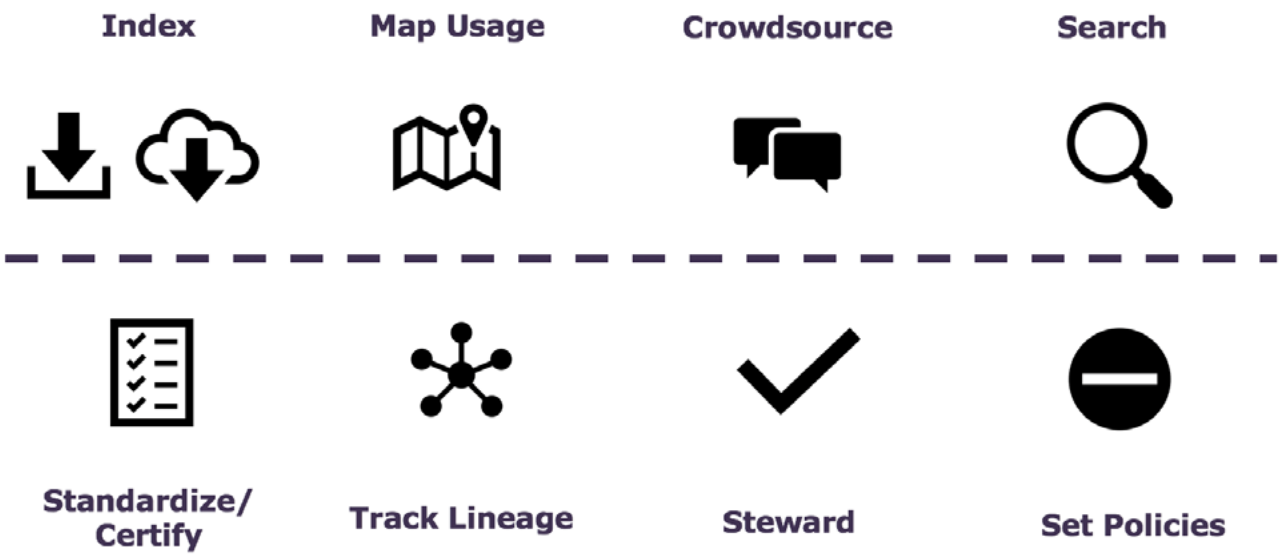
Catalogs have also evolved to support cloud deployments and public cloud services. This is crucial. Cloud-based data lakes are proliferating as companies take advantage of the massive-yet-elastic computing power, low-cost object storage, and advanced ML and AI services available in public clouds.

Today's most comprehensive data catalogs address the gambit of data sources, whether on-premises or in the cloud, and they also address data-related assets, including refined data sets, reports, schema and even models and APIs.

As shown in Figure 3, best-of-breed catalogs make all of the sources and assets described above more visible, understandable and securely accessible with the following capabilities:

- **Index.** Data-crawling capabilities connect to data sources, read metadata and automatically populate the catalog, indexing data and related assets by source, format, physical location, creation date, definition and other attributes. Best-of-breed products also include a business glossary that enables organizations to label and describe data in the language of the business and then link these terms to the technical metadata in source systems.

Figure 3. Data Cataloging and Data Governance Capabilities



Source: Constellation Research

- **Map usage.** This deeper, typically ML-powered capability logs and maps data inputs, outputs, and access and query patterns. Benefits include enriched tagging and statistical understanding of how data is used and by whom and how data assets are related. Advanced ML functionality can also generate recommendations based on user- and role-based behavior patterns.

- **Crowdsource.** Collaboration features are crucial to an effective data catalog. Leading products facilitate stakeholder communication around data assets and related projects and workflows. Advanced social capabilities capture "tribal knowledge" in the form of annotations, ratings and recommendations.

- **Search.** The key benefit of the catalog is making data and related assets visible to both business users and more technically savvy constituents through a user-friendly search interface. Advanced catalogs support natural-language querying, going beyond keywords to discern the intent of phrases and questions. Some catalogs capture samples of data so users can physically see and better understand exactly what data sets contain. Shopping-cart-style interfaces are used in some products to support actions such as data-access requests.

Constellation's POV: Data catalogs are proliferating, but be aware of that some catalogs offer only a platform- or cloud-centric view of data. Apache Ranger, for example, is geared to cataloging data in Hadoop clusters. Similarly, the AWS Glue Data Catalog is only natively integrated with Amazon cloud services such as EMR, Athena and Redshift Spectrum. A comprehensive data catalog synchronizes with these and other single-platform data catalogs while also indexing and profiling other sources and assets to provide an enterprisewide view of data and assets. This gives catalog users a complete view of data and assets available across the organization, regardless of location or deployment choice.

Vendors delivering or developing comprehensive data catalogs aimed at providing an enterprisewide view of data and assets include Alation, Cambridge Semantics, Collibra, IBM, Infogix, Informatica, Microsoft, Oracle, Qlik (through its recent acquisition of Podium Data), Reltio, Talend, Unifi Software and Waterline Data.

The next question is to what degree a catalog fulfills or integrates with the data-governance functions shown below the dotted line in Figure 3. At a minimum, catalogs should include a business glossary, which eases the challenge of finding data by supporting consistent, business-user-friendly search terms. A glossary shields non-technical users from the need to know or interpret arcane data platform terminology.

But Constellation's view is that catalogs should go beyond indexing and glossary functionality. For starters, tracking of data lineage is a governance function that should be supported by a comprehensive data catalog. But as discussed below, lineage addresses only part of the total data-governance challenge. Data standards, certifications and usage policies are all aspects of governance that bring additional context to data. If the role of the catalog is to not just connect users with data but to improve their understanding of that data, then integration with data-governance capabilities is crucial.

Data Governance Ensures Consistency and Trust, Sparks Innovation

Data inconsistencies, data quality problems and lapses in proper usage are data governance issues that, when not addressed, have a corrosive effect on trust. And when trust is diminished, it undermines the value of the data and faith in the data team.

Here's a closer look at more advanced functionality to look for from deeper data catalogs and integrated data-governance platforms.

- **Standardize/certify.** Comprehensive solutions go beyond glossary functionality to address metadata management and standardization of terms (also known as reference data). Catalogs should be able to synchronize with data-certification capabilities tied to data-integration and analytics platforms, giving greater weight to authoritative sources and official, watermarked or approved reports. Some catalogs are built on master data management (MDM) platforms, but MDM products tend to be designed for information management professionals and data stewards, so consider ease of use for business users when assessing such products. Comprehensive governance solutions also go beyond crowdsourced thumbs-up/thumbs-down ratings, supporting customizable workflows and standards for measures of data completeness and ratings for data quality.

- **Track lineage.** Knowing the provenance of data is a must in any regulated environment. The tracking task gets more challenging as data pipeline complexity mounts and includes the cloud. Ensure that data-governance functionality spans data platforms, data-integration and data-transformation options, and both on-premises and cloud-based deployments. Regulations are often country specific, so it's also crucial to know exactly where data is processed and stored at all points along a data flow.

- **Steward.** Data catalogs support data curation, but they don't tend to include data-stewardship functionality, such as role-based, customizable workflows that connect stakeholders including stewards, subject-matter experts, data owners and process owners. Products with advanced integration features support embedding of stewardship tasks into data management, MDM, and data-quality processes. Leading products also include automation and integration with data policy management features (see below), which expedite and scale up the execution of requests for data access, the enforcement of data-sharing agreements and other actions.

- **Set policies.** Comprehensive data governance includes policy management capabilities that enable data stewards, owners and compliance teams to detail security, legal and strategic requirements. Policies might advise would-be users on best-use cases for data, the anticipated lifespan of a data source, and terms of data access and use. Ideally, policy management capabilities are or can be integrated with the search and shopping-cart features of the catalog. This brings together data-access and data-request workflows with the enforcement of access controls, data-sharing agreements, and encryption and masking requirements.

Constellation's POV: Providing access to data is just the first step. Mature organizations recognize that data is only as valuable as it is trusted. To be trusted, data must be made accessible in accordance with required data policies and procedures. That's why Constellation believes data catalogs need to be tightly integrated with data governance capabilities, including metadata management, data-lineage tracking, data stewardship and data policy management capabilities. When better data access and trust can be combined with effective governance and enforcement of policies, organizations gain a deeper

understanding of data. This understanding, in turn, leads to greater confidence to use data in new and innovative ways, potentially driving new data-driven services and business models.

RECOMMENDATIONS

Based on Constellation's conversations with clients, surveys, vendor input and best practices, Constellation recommends the following:

- **Provide self-service data access tied to governance capabilities.** Catalogs ease access to and understanding of data, but they must also integrate with governance capabilities. Sophisticated catalogs can support data policies and classifications, such as "public," "regulated" and "restricted." At the same time, that can encourage innovation by revealing the existence and descriptions of data (to support requests for data access and potential new uses of data) without exposing the actual content.

"A comprehensive data catalog indexes the content of data platforms, schema, models, APIs, analytic data sets and reports, whether deployed in the cloud or on-premises."

Constellation's POV: Data catalogs are proliferating, but some address only one platform or deployment environment. A comprehensive, enterprise data catalog indexes the content of data platforms, schema, models, APIs, analytic data sets and reports, whether deployed in the cloud or on-premises.

- **Ensure well-defined data and business-user-friendly terms.** Data catalogs often include business glossaries for defining user-friendly data descriptions, but make sure that the catalog can link business terms to technical metadata.

Constellation's POV: Seek out more-sophisticated products that support the standardization of reference data to help avoid inconsistencies and redundancies that diminish data quality and trust.

- **Deliver a platform for data collaboration.** Data catalogs should include collaborative capabilities so stakeholders ranging from business users and data owners to data stewards, engineers, analysts and data scientists can communicate around data and related assets, projects and initiatives.

  Constellation's POV: General-purpose collaborative platforms are a better option than email for data collaboration, but data catalogs do a better job of tying comments and communications to specific data, assets and projects without relying on collaborators to document the context of their exchanges.

- **Capture tribal knowledge on data quality and suitability.** Best-of-breed data catalogs go beyond communication to include functionality for capturing source-data- and asset-specific annotations, ratings and reviews.

  Constellation's POV: Seek out products that support configurable workflows and measures for detailing data completeness and data-quality scores.

- **Ensure support for managing data policies and data-sharing agreements.** Look for data-governance mechanisms for establishing and enforcing data policies and data-sharing agreements. Policies might cover whether and how data can be shared internally and externally. Data-sharing agreements might include expiration dates and requirements for encrypting and masking data.

  Constellation's POV: Support for data-policy management and data-sharing agreements are must-haves for regulated organizations and any business interested in sharing data with third parties or monetizing information.

- **Start small and ensure business participation.** Big-bang deployments of data catalogs and data-governance capabilities rarely succeed. You'll get faster, better results by starting in targeted areas in which business stakeholders are willing participants. Build on success by moving into adjacent data domains.

  Constellation's POV: Devote time to training people on roles, such as data owner or data steward. Promote the benefits of trusted data to ensure adoption.

- **Get help from systems integrators.** Data-catalog and data-governance platform vendors and systems integrators offer invaluable experience and guidance on implementing technologies and establishing policies and procedures for data access and data governance. Systems integration firms including Accenture, Capgemini, Cognizant, Infosys, TCS and Wipro and management consulting firms such as First San Francisco Partners, Information Assets, Kalypso and Slalom can help with the crucial people and process sides of data-governance programs.

  Constellation's POV: Data catalogs and data-governance technologies can certainly help with providing agile access to trusted data, but you can't just buy good governance and compliance. Solid processes and workflows and thorough user training and follow-up are crucial to ensuring user adoption and successful data-governance initiatives.

ENDNOTES

[1] *Chris Kanaracus, Courtney Sato and R "Ray" Wang, "Constellation Research 2017 Digital Transformation Study," Constellation Research, October 2017. N = 105 CIOs and senior line-of-business and IT executives.* **https://www.constellationr.com/research/constellation-research-2017-digital-transformation-study**

[2] *Dion Hinchcliffe, "Agile IT—2017 CIO Survey," Constellation Research, April 2017. N = 51 CIOs, CTOs, CDOs and VPs of IT at innovative organizations.*

[3] *"Constellation Research 2017 Digital Transformation Study."*

Doug Henschen

Vice President and Principal Analyst

Doug Henschen is Vice President and Principal Analyst at Constellation Research, Inc., focusing on data-driven decision making. His Data-to-Decisions research examines how organizations employ data analysis to reimagine their business models and gain a deeper understanding of their customers. Data insights also figure into tech optimization and innovation in human-to-machine and machine-to-machine business processes in manufacturing, retailing and services industries.

Henschen's research acknowledges the fact that innovative applications of data analysis require a multi-disciplinary approach, starting with information and orchestration technologies, continuing through business intelligence, data visualization, and analytics, and moving into NoSQL and big data analysis, third-party data enrichment, and decision management technologies. Insight-driven business models and innovations are of interest to the entire C-suite.

Previously, Henschen led analytics, big data, business intelligence, optimization, and smart applications research and news coverage at *InformationWeek*. His experiences include leadership in analytics, business intelligence, database, data warehousing, and decision-support research and analysis for *Intelligent Enterprise*. Further, Henschen led business process management and enterprise content management research and analysis at *Transform* magazine. At *DM News*, he led the coverage of database marketing and digital marketing trends and news.

@DHenschen **constellationr.com/users/doug-henschen** **linkedin.com/in/doughenschen**

ABOUT CONSTELLATION RESEARCH

Constellation Research is an award-winning, Silicon Valley-based research and advisory firm that helps organizations navigate the challenges of digital disruption through business models transformation and the judicious application of disruptive technologies. Unlike the legacy analyst firms, Constellation Research is disrupting how research is accessed, what topics are covered and how clients can partner with a research firm to achieve success. Over 350 clients have joined from an ecosystem of buyers, partners, solution providers, C-suite, boards of directors and vendor clients. Our mission is to identify, validate and share insights with our clients.

Organizational Highlights

- Named Institute of Industry Analyst Relations (IIAR) New Analyst Firm of the Year in 2011 and #1 Independent Analyst Firm for 2014 and 2015.
- Experienced research team with an average of 25 years of practitioner, management and industry experience.
- Organizers of the Constellation Connected Enterprise—an innovation summit and best practices knowledge-sharing retreat for business leaders.
- Founders of Constellation Executive Network, a membership organization for digital leaders seeking to learn from market leaders and fast followers.

www.ConstellationR.com

@ConstellationR

info@ConstellationR.com

sales@ConstellationR.com

San Francisco | Belfast | Boston | Colorado Springs | Cupertino | Denver | London | New York | Northern Virginia
Palo Alto | Pune | Sacramento | Santa Monica | Sydney | Toronto | Washington, D.C.