Talena™

WHITEPAPER

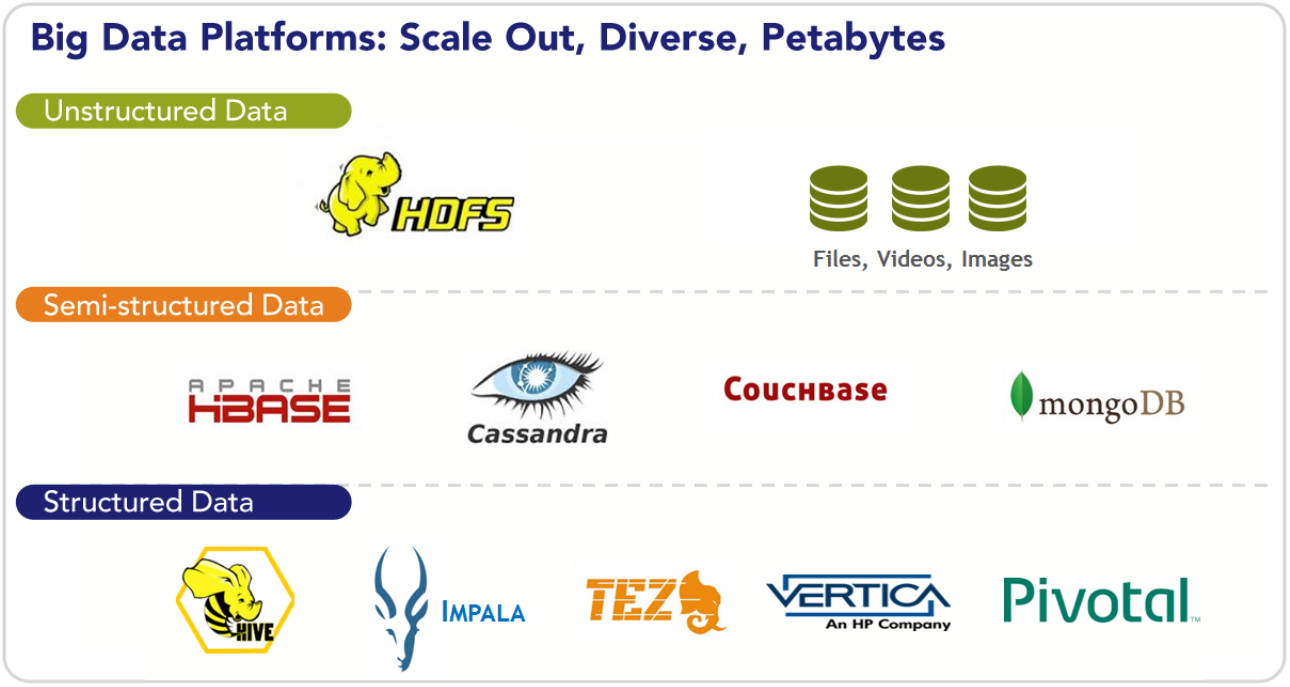
A Technical Perspective on the Talena Data Availability Management Solution

BIG DATA TECHNOLOGY LANDSCAPE

Over the past decade, the emergence of social media, mobile, and cloud technologies has created vast quantities of data, or "Big Data", and its associated storage. More recent developments in the Internet of Things, and the widespread adoption of these platforms, have accelerated the Big Data growth to new levels.

Early technology adopters such as Yahoo!, Google, and Facebook had to find economical ways to store and process this Big Data. Commercial third party technologies were not an option since they are very expensive and not designed to handle the growth and scale of Big Data. Consequently, these companies resorted to developing in-house solutions that could scale-up to handle the data volumes and growth, and leveraged industry-standard hardware to align with open source economics. Eventually, they open-sourced their software to make it available to other companies facing similar challenges.

The Big Data technologies developed fall into the three main categories shown below.



Over the last few years, these technologies have matured, resulting in widespread adoption by organizations that needed to store and process large quantities of data. To extract value from this data, organizations created new applications that are increasingly at the center of business operations. Given their importance in data analysis and decision making, these applications are being elevated to Tier 1 status.

Operational Challenges with Big Data

Talking to hundreds of big data users revealed that making copies of production data is a common task performed by engineers and DevOps teams. Copies are required for test, development or analytics purposes, as backups to protect against user error or application corruption, for disaster recovery purposes, or simply to archive and retain old data. However, making copies of tens or hundreds of terabytes poses some unique challenges.

1. Copying petabytes of data multiple times (once for each use case) is not practical and negatively impacts production applications and networks.
2. Storing multiple copies of data for each use case quickly multiplies the costs associated with a big data deployment.
3. Typically, a dedicated team of engineers writes, customizes, maintains, and runs scripts as needed. These scripts are non-trivial and require constant maintenance and updates. Additionally, most data copy requests are custom in nature and involve writing new scripts each time. Most organizations would rather use their scarce engineering resources for other more valuable tasks.

IT operations or DevOps teams responsible for the big data infrastructure face two operational challenges.

- To ensure recoverability from user errors, application corruption, and disasters, they must write scripts to ensure the data is adequately protected.
- If 30% annual data growth is not proactively managed, a 100 node cluster can quickly grow to 220 nodes resulting in a proportional increase in both capital and operational expense.

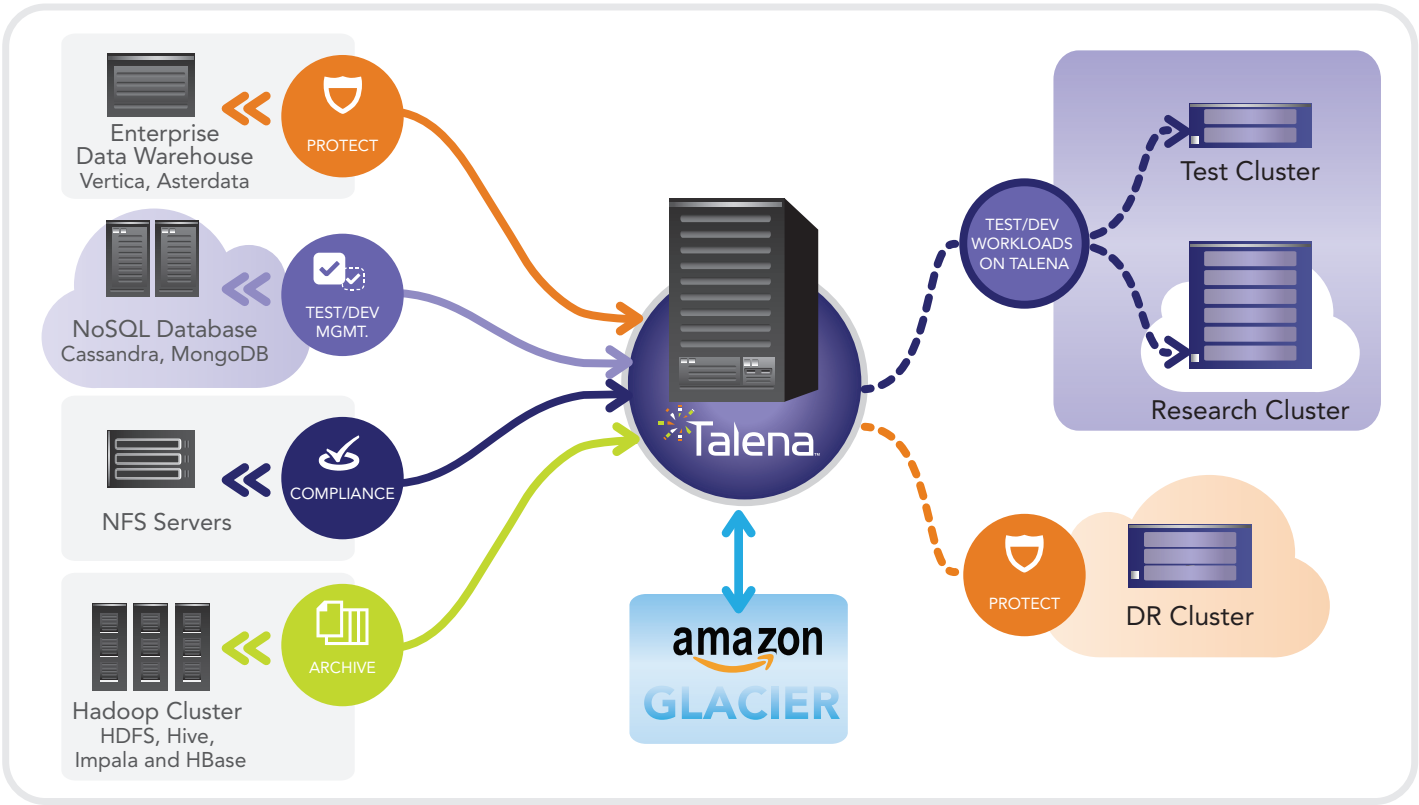
Engineering teams responsible for managing test and development environments also face operational challenges.

- They must use scarce engineering resources to write scripts to copy data to new locations.
- Script writing and data copies take time. Data consumers (data scientists, developers, testers) must wait for the data to be made available, resulting in lost productivity.

Talena Product Description

Talena is an enterprise software product that addresses the operational challenges associated with managing big data environments like Cassandra, HP Vertica, Impala, and Spark. Talena's philosophy and architecture is to make only one additional copy of data, and then use that copy for multiple purposes such as testing, development, backup, and disaster recovery. This significantly reduces both the impact on production networks and storage infrastructure costs. Talena software eliminates all scripting by providing customizable workflows and policies that automate the entire process.

Talena software runs on industry standard x86 hardware — on a physical server, inside a virtual machine, or in a cloud environment. Each system running the Talena software is called a node. A collection of nodes is called a Talena cluster and is a highly available, scalable, and distributed system. The Talena software creates a single large storage pool using direct-attached storage in the nodes. The single storage pool is used to store the data managed by Talena. Talena software also ensures the cluster, and the data stored on the cluster, is always available in case of cluster hardware failures.



When deployed, the Talena cluster can be in the same data center as your production Big Data repositories, or it can be in a different data center, or even in the cloud. The Talena cluster then interacts with the various Big Data repositories (both production and non-production) to automate, manage, and optimize the data copy process.

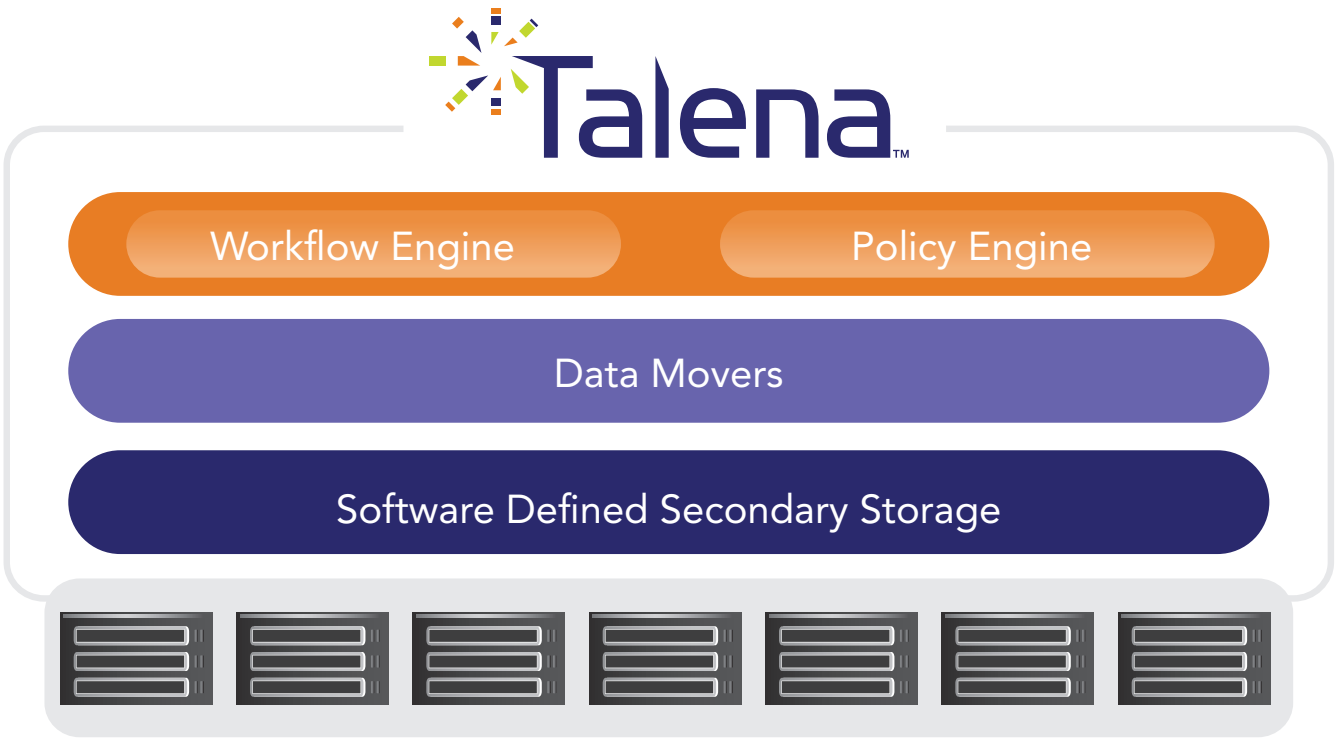
Using the web-based Talena user interface, administrators can define flexible workflows that automate the process of periodically extracting the relevant data from various big data repositories and copying it to a specified location, such as the Talena cluster itself, Amazon Glacier, or a test Hadoop cluster.

Talena Software Architecture

The Talena software architecture has three main components:

- Software Defined Secondary Storage
- Data Movers
- Workflow and Policy Engines

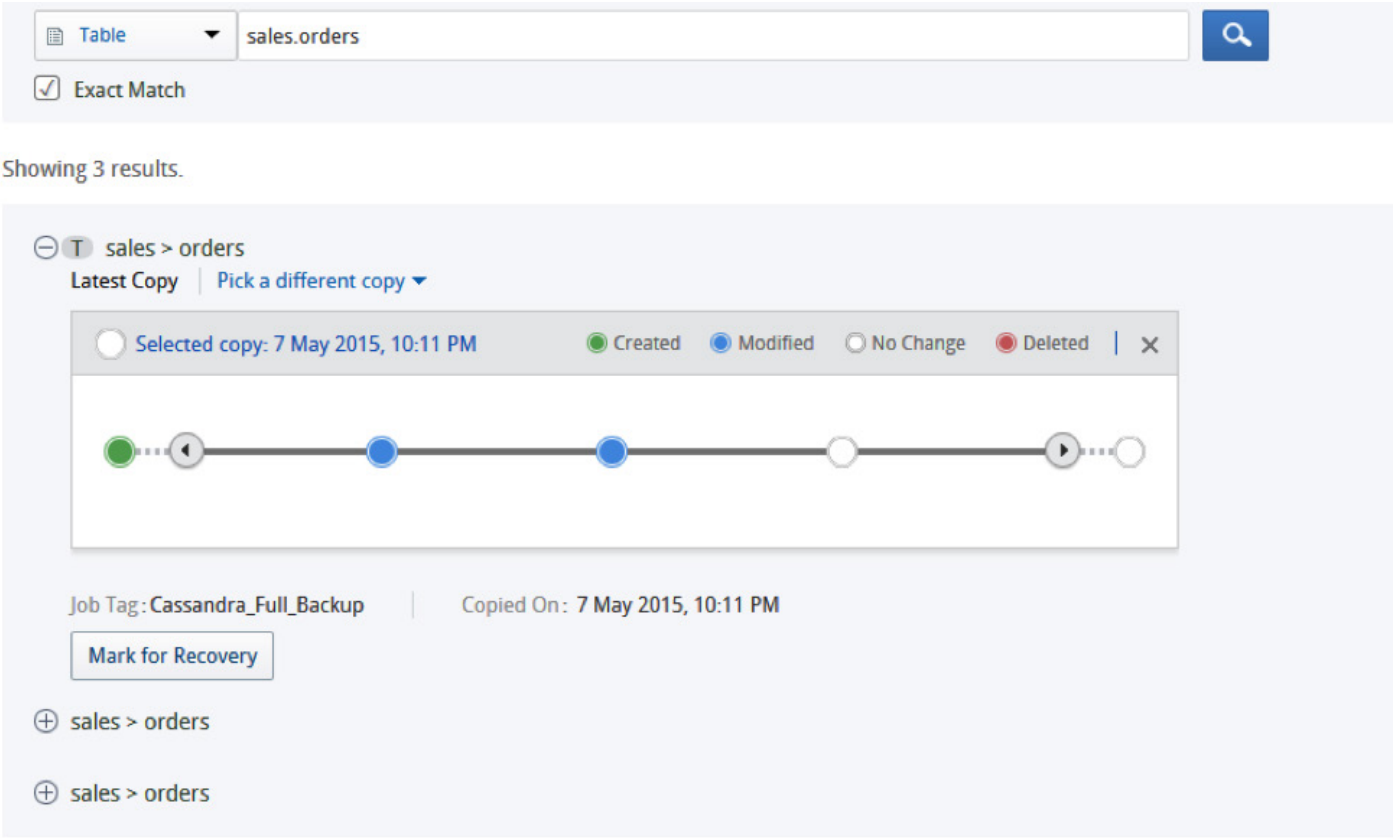
The relationship and details of these components is shown below.



1. Software-defined Secondary Storage

The software-defined secondary storage component ensures that data is stored efficiently and reliably on the x86 server's available internal storage. To accomplish this, the Talena software incorporates a distributed file system that creates a large storage pool from the internal disk drives of the various x86 servers. To optimize the storage footprint, Talena developed software with distributed block-level deduplication capabilities employing aggressive compression algorithms. The Talena file system also supports snapshots that enable storage of multiple point-in-time backup copies with minimal storage overhead. Finally, to ensure data resiliency, the software uses erasure coding to survive multiple node failures. To ensure fast and easy data retrieval, Talena provides a distributed metadata catalog - Talena FastFind™ - for all data stored on the Talena file system.

With Talena FastFind, users can quickly search and retrieve the relevant data of interest. The screenshot below shows the recovery process, and demonstrates the power and ease-of-use associated with searching for a database table and doing a point-in-time table recovery from a previous backup.

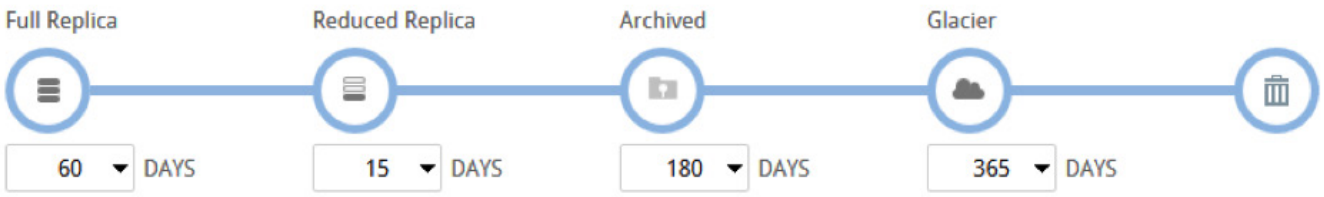


2. Data Movers

Application specific data movers interact with the various Big Data repositories that Talena supports and requires no software installation on production nodes. The data movers understand the nuances of the different repositories and provide the mechanism to pull out, or restore, data, to the repository in a consistent manner. For example, when working with Hive, the Hive data mover presents the Hive schema to the user (tables and databases vs. files and directories), and backs up and recovers the Hive metastore. The data movers are very efficient and only move data that has changed since the last time data was captured (incremental-only). Additionally, data movers use a lot of parallelism so that performance scales as the size of the Talena cluster grows.

3. Workflow & Policy Engine

Workflows and policies make Talena software extremely flexible. Workflows can be easily configured for different use cases such as backup, archives, data copies for test and development purposes, and disaster recovery. When making copies of production data for test and development purposes, the data may also be downsized by specifying a sample size and anonymizing certain attributes by specifying masking parameters.



Policies are also applied to a workflow to define the frequency of execution, the priority of the workflow, and use case specific attributes such as archiving criteria, and backup retention periods.

Workflows are recovery-centric and provide a lot of power and flexibility. In addition to enabling fast search and recovery using Talena FastFind, recovery-centric workflows allow objects to be renamed on recovery and enable restoration to alternate clusters of different size and configuration. Very granular point-in-time recoveries are also possible using Talena FastFind.

TALENA USE CASES

The Talena software can be accessed using a web-based user interface and can be configured to support a number of different use cases such as:

Self-service access to production data for non-production use:

Data science and engineering teams often require sandboxes for analysis, test, development, and other activities. These sandboxes are populated with production data subsets after obfuscating Personally Identifiable Information (PII) and other confidential information. This is usually accomplished by scarce engineering resources writing custom scripts, taking these expensive resources away from their real jobs. Further, data scientists must wait for the scripts to be written and executed, resulting in productivity losses. By enabling a complete self-service model, Talena software eliminates scripting development delays by allowing data scientists and administrators to set up automated pipelines to copy data to non-production sandboxes on a periodic or ad hoc basis. Talena ensures that sandbox data is statistically consistent and does not lose relationships present in the original data. Additionally, PII is masked in a consistent manner and data is transformed as needed. No time is lost waiting on data and scarce data engineering resources that are better used for more important tasks.

CASE STUDY:
A Fortune 100 technology company deployed in production a 500 TB, 200-node Hadoop cluster. With Talena, the company was able to automate the entire test and development, and backup process, for its Hadoop cluster at a fraction of the cost of using their engineering resources to write and maintain custom scripts.

Protecting valuable data from user error, corruption and disasters:

Many enterprises run Hive, Spark, Impala, or Tez to replace their traditional Enterprise Data Warehouse (EDW). NoSQL databases such as Cassandra and HBase, and newer EDWs such as Vertica are often used in critical large scale enterprise applications. As these applications become Tier 1, enterprise IT operations groups often look for industrial grade backup solutions similar to those existing for legacy platforms. Talena provides application-aware incremental-forever backups for these large distributed databases with block-level de-duplication, granular recoveries, and fast metadata search with Talena FastFind. Talena's disaster recovery solution provides a single platform for multiple data sources.

By using techniques such as block-level deduplication and deep compression, the disaster recovery footprint can be up to 5X smaller. Finally, unlike traditional DR solutions where the DR resources (compute and data) are unusable, with Talena, the DR resources can be used to run secondary workloads.

CASE STUDY
One of the world's largest stock exchanges transitioned from an older data warehouse to one built on top of Impala and Hadoop. With Talena, the company was able to deploy an enterprise grade backup and recovery solution to protect against user errors, and application corruption or site disasters.

Containing the growth of production clusters by archiving old data:

Data typically grows faster than an organization's IT budget. Talena offers a policy-driven solution that automates the archival of older, less-frequently-used data to different storage tiers including cloud storage (Amazon Glacier). By using techniques such as block-level de-duplication and deep compression, Talena allows archived data to be stored very efficiently with up to an 80% reduction in storage footprint. The archived data can be actively used to run secondary workloads such as queries against historical data. Compliance and retention can be enforced through WORM policies that prevent modifications or deletion of files. Data durability is critical for archival and retention and is accomplished by using techniques such as erasure coding and SMART disk diagnostics. To secure sensitive data, Talena incorporates features such as data masking, role-based access controls, LDAP authentication, and Kerberos support.

CASE STUDY
A publicly traded ad-tech company needed an archive solution for its growing 900-node Hadoop cluster. With the policy-driven and automated Talena solution, the company was able to significantly lower costs by moving old data to Amazon Glacier and still enjoyed the added flexibility of a built-in catalog that facilitated rapid searches using Talena FastFind.

SUMMARY

With the rapid adoption of Big Data technologies such as Cassandra, Impala, Hive, and Vertica organizations are quickly realizing the need for a comprehensive, robust, and enterprise grade data availability management solution similar to those available today for legacy applications. Talena data availability management software allows organizations to simply and efficiently manage data in Big Data repositories for many purposes, including test and development, management, backup and recovery, archiving, or disaster recovery. Many organizations have realized tremendous benefits by replacing current homegrown solutions with Talena software. By deploying Talena software organizations they are able to improve application recoverability, eliminate sensitive production data from being copied to test and development environments, and to significantly reduce the costs of managing their Big Data infrastructure.

About Talena

Talena provides the first data availability management software expressly engineered for Big Data technologies, enabling always-on data across the full lifecycle of applications so they can meet internal and external service level agreements. Our data availability management software is uniquely designed to optimize the test/dev management, backup, recovery, and archive functions so that our customers can iterate rapidly on their applications, prevent data loss, and minimize compliance risk. Based in Silicon Valley, we are backed by Canaan Partners, Intel Capital, ONSET Ventures, and Wipro Ventures.

For more information, visit www.talena-inc.com.

Please contact us for more information at info@talena-inc.com or visit us at www.talena-inc.com.

Talena, Inc. | 2860 Zanker Road,, Suite 109, San Jose CA 95134 | +1.408.649.6338 | talena-inc.com

© 2016 Talena, Inc. All rights reserved. Talena and the Talena logo are trademarks of Talena in the US and in other countries. Information subject to change without notice. All other trademarks and service marks are property of their respective owners.