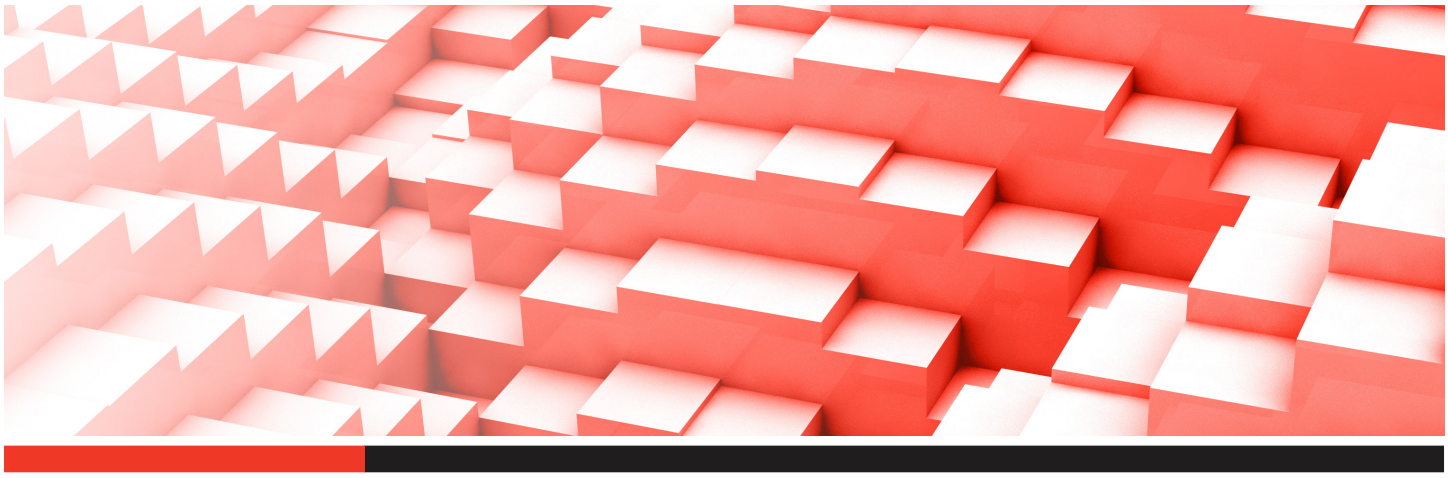
IMANIS DATA



WHITEPAPER

An Inside Look at the Imanis Data Architecture

INTRODUCTION

Big Data platforms including Hadoop and NoSQL have matured beyond the confines of the research lab, and now serve as the foundation for high-value, business-oriented applications. Yet while most enterprises have happily embraced specialized Big Data technologies for gathering, organizing, and extracting meaning from this information, far too many continue to rely on haphazard and potentially risky methods for safeguarding these assets and dispensing them to software developers and testers.

The Imanis Data Management Platform closes this gap, delivering affordable, flexible, enterprise data management for these new mission-critical applications. This paper describes the challenges of modern data management, explains why an updated approach is needed, and supplies technical details about the Imanis Data solution.

The intended audience includes:

- Database administrators
- Architects
- IT operations
- Software developers
- Quality assurance professionals
- DevOps teams

ENTERPRISE DATA MANAGEMENT REQUIRES A FRESH APPROACH

Transactional software solutions such as Customer Relationship Management (CRM) and Enterprise Resource Planning (ERP) used to be the primary sources of business information and were typically hosted in relational databases. This is no longer the case: a diverse and ever-expanding assortment of new technologies is spawning enormous amounts of raw data that are driving entirely new types of applications.

In an effort to keep pace with the volume, variety, and velocity of all this supplemental information – collectively labeled as Big Data – businesses are rolling out a diversified collection of applications built on scale-out file systems and databases such as:

Hadoop-related ecosystem

- Hadoop File System (HDFS)
- Hive
- HBase
- Impala
- Tez
- Spark

Massively Parallel Processing (MPP) data warehouses

- Vertica
- IBM Netezza

NoSQL databases

- Cassandra
- Couchbase
- HBase
- MongoDB

Until fairly recently, most businesses primarily employed these Big Data platforms for research or proof-of-concept projects. This meant that secure information access and robust data protection were relatively minor considerations. However, everything changes once an organization begins to merge these assets into their core application portfolio: it's no longer acceptable to risk a system outage, data loss, or security breach.

Although the majority of organizations have longstanding tools and techniques for protecting traditional data sets as well as sharing it with software developers and testers, there are significant challenges where Big Data is concerned.

Data protection issues

Traditional backup and recovery techniques - and supporting products - remain ideal for their original purposes, but are a mismatch for the Big Data technologies listed earlier. This gap leaves organizations exposed to data loss, downtime, and cyber attacks

Many of these new platforms supply built-in replication, which entails distributing multiple data copies onto distributed servers. While this approach reduces the possibility of outright data loss, it paradoxically serves to propagate user or application-driven data corruption: damage quickly spreads to all replicated copies. This means that it's also essential to implement a proper backup/recovery strategy for Big Data.

While the majority of these new platforms ship with their own backup and restore utilities, these are laden with substantial drawbacks:

- They're myopically focused on a single technology, and thus don't recognize the heterogeneous Big Data portfolio that prevails in many enterprises
- They're driven by command line interfaces (CLI), which demands manual interaction, making both backup and restore processes cumbersome and error-prone
- They lack automation, a fundamental characteristic of a dependable data protection process.
- They're restrictive in terms of their capabilities, thus minimizing their effectiveness as companies scale their data infrastructure

Scripting is a particularly inefficient way to protect data: it siphons off valuable IT talent from mainline business responsibilities, and produces a continually growing inventory of brittle, maintenance-intensive assets.

Data mobility challenges

To produce effective applications, software developers and testers need prompt access to meaningful amounts of representative production information from Big Data repositories. Faced with these requirements, enterprises have generally resorted to one of two approaches: create fabricated test data, or write scripts to transfer data from production to development and testing environments.

Organizations that rely on artificially constructed data confront the risk of delivering solutions that are divorced from reality, and are prone to software quality issues once they are placed into production and encounter authentic information.

For those enterprises that do permit information extraction from production Big Data systems in support of new software development and testing, the task of writing extraction scripts hamstrings and delays the entire new application creation process. In fact, according to a comprehensive study commissioned by Imanis Data, 90% of organizations defer application rollouts waiting for data. What's more worrisome – and possibly exposes the business to legal consequences – is the potential for inadvertently divulging highly sensitive Personally Identifiable Information (PII) to unauthorized users.

Given these data protection and mobility deficiencies, it's clear that a fresh, comprehensive approach is necessary to support the Big Data platforms that are now core enterprise resources. Imanis Data provides the first data management software solution expressly designed for modern data platforms, helping companies protect valuable data assets and iterate rapidly on their business-critical applications through capabilities such as backup & recovery, test/dev management, and archiving.

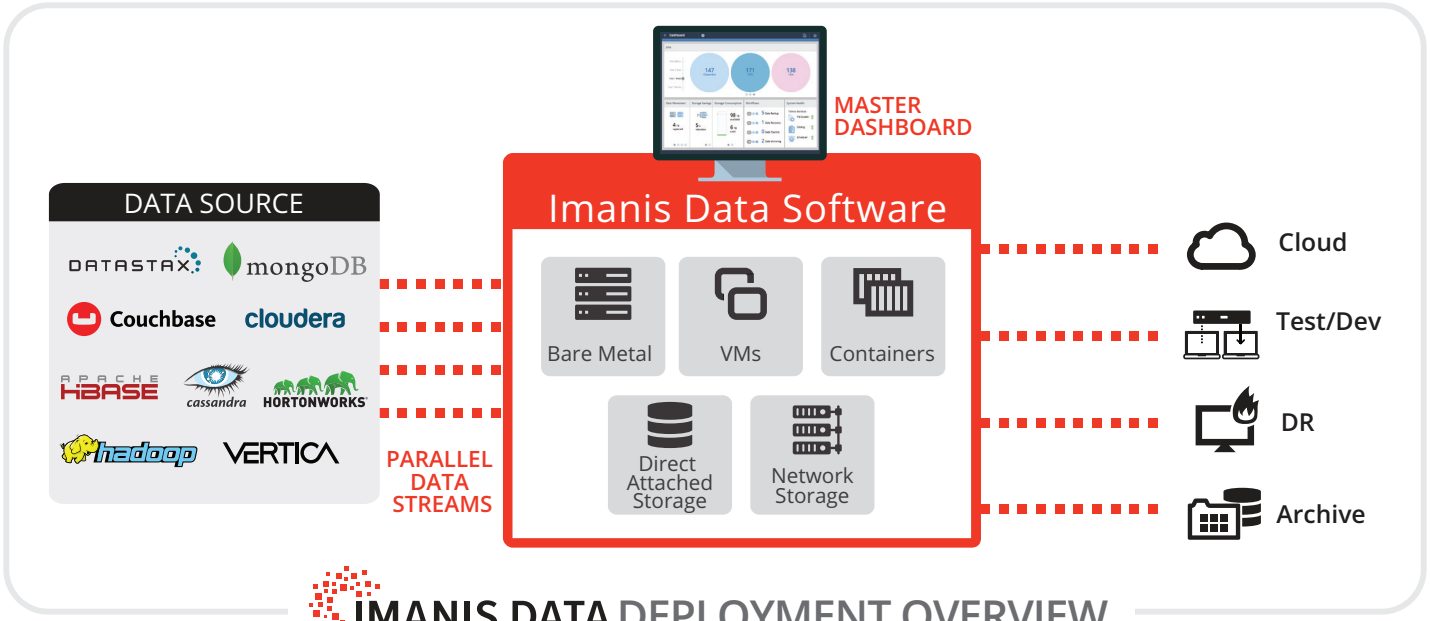
The next section describes Imanis Data's technical architecture and merits.

INTRODUCING THE IMANIS DATA PLATFORM

Recognizing the inadequacy of existing data management methods Imanis Data has created a unique new approach based upon three design pillars:

1. Petabyte scale
2. Data awareness
3. Machine learning

The following sections describe how the Imanis Data architecture attained these objectives.



IMANIS DATA DEPLOYMENT OVERVIEW

1. Petabyte scale

Software-Defined with Massive Scalability

Since scale-out – which leverages the power of commodity servers to increase capacity – is a fundamental Big Data architectural philosophy, it's natural that the same must hold true for the technologies that are tasked with protecting and managing this information.

Each software-defined, storage-agnostic Imanis Data node is deployed on cost-effective commodity hardware, using directly attached disk drives to maintain backed-up information. This eliminates any reliance on more expensive Network Attached Storage (NAS) or Storage Area Network (SAN) components, although customers are welcome to make use of these devices if desired. It also frees Imanis Data from mandating a specific hardware configuration.

All nodes in the Imanis Data cluster establish connections to one or more nodes in the primary cluster, so that data can be transferred concurrently. Adding more nodes in Imanis Data increases the data ingestion rate from the production cluster, which is a completely parallelized and highly scalable technique. Horizontal scalability is also a major attribute of Imanis Data's catalog, which is a cornerstone of data restoration. The catalog is capable of tracking and versioning millions of objects, and offers full search capabilities.

Imanis Data's decision to avoid any hardware mandates – such as proprietary, expensive appliances also resulted in numerous cost saving benefits:

- It can be installed on bare metal servers or virtual machines
- It's available on popular enterprise Linux distributions:
  - RHEL
  - Ubuntu
  - Oracle Linux
  - Centos
- It utilizes inexpensive, commodity disk drives
- It's capable of running in the cloud, on-premise, or any combination
- It's able to be instantiated with a small number of servers, and automatically rebalances its clusters as more servers are added

- It provides administrators with a "single pane of glass" that lets them support multiple data sources and use cases, as well as manage all aspects of their Imanis Data environment
- It integrates with Nagios agents that transmit alerts regarding disk space, backup failures, and other issues, along with daily email notifications about job status

Finally, many customers elect to deploy Imanis Data using the same operating system as their production data nodes. This homogenizes their operating system-level software infrastructure, helping to reduce maintenance efforts and potential security issues.

Agentless architecture

Customarily, enterprise backup solutions have mandated installing software agents on all data nodes. These components are then tasked with the job of transferring information. This tactic simply won't work in ever-changing Big Data environments: administrators would quickly be overwhelmed with maintenance and monitoring responsibilities, and these far-flung agents would also introduce inherent security risks.

Instead of forcing customers to install agents throughout their production landscape, Imanis Data's agentless architecture uses the already-optimized public interfaces supplied by the Big Data platform vendors as its cross-system communication pipeline.

Storage efficiency

Given how large Big Data environments can get, it's no surprise that storage expenditures can quickly outpace expectations, especially when including the outlays that enable sufficient information backups. To help keep this overhead to a minimum, Imanis Data employs multi-state data reduction techniques.

Imanis Data's global, data-aware, variable-length deduplication engine is the initial component in this workflow. It begins the storage reduction process by identifying the data that is to be deduplicated. This information may be stored in one of many formats, such as compressed files (e.g. GZ, Snappy, LZO, and so on) or application-specific structures (e.g. RCFile, Parquet and ORC for Hive and Impala or SSTable for Cassandra). Once deduplicated and compressed, the output is saved onto the file system and erasure coded to increase its durability.

The upshot is that Imanis Data offers significant storage savings that increase as the volume of data increases.

Incremental-only backups

For many years, the process of backing up enterprise data has been comprised of a full weekly backup, followed by daily incremental backups. This time-tested approach falls flat in Big Data environments that are commonly measured in petabytes: it's simply too hardware-intensive, laborious, time-consuming, and error prone to conduct weekly full backups.

Imanis Data's incremental-only backup strategy is much more appropriate for Big Data technologies. It uses snapshots to ascertain what's changed since the previous incremental backup, and then only backs up relevant information. Incremental data is immediately materialized onto Imanis Data servers, and a restore point is created. This fully materialized restore point – driven by the incremental backups – can be used to return data to the production clusters as-is, without requiring any additional work.

Robust security

Imanis Data implements industrial-strength access, security, and encryption capabilities, including:

- Kerberos and LDAP authentication
- User and role management
- Support for encrypted file systems
- SSL for data transmission between Imanis Data and primary clusters

2. Data Awareness

Being data aware means understanding not only the data being managed but also the schema in which it is stored (e.g. MongoDB, Hadoop). This enables Imanis Data to more intelligently protect, orchestrate, and automate all data management tasks. Additionally, data awareness means adapting to the dynamic nature of Big Data environments. It's highly unusual to encounter a static, homogenous Big Data implementation. Instead, most organizations deploy a unique blend of tools and technologies that were meticulously selected to meet their particular needs. Furthermore, these configurations are constantly changing by incorporating additional Big Data platforms and adding, adjusting, or dropping data nodes. In addition, Imanis Data was engineered to thrive in every type of landscape, with an intelligent set of data-aware features.

Heterogeneous Big Data platform support

Each Big Data product offers highly targeted capabilities that enterprises earmark for achieving specific goals. For example, the use case for a Hadoop installation is very different than for a MongoDB instance. This means that it's likely that a business will acquire an assortment of these platforms to thoroughly address their Big Data needs.

Consequently, any product that's intended to strengthen an organization's collective Big Data implementation must address all of the deployed technologies. Imanis Data attains this objective by blending a loosely-coupled core architecture, deep insight into each supported product, and platform-specific data movers that completely utilize each of the vendor-supplied APIs. This design principle also makes it straightforward for Imanis Data to integrate new Big Data solutions based on customer requirements.

Topology independent backup and restore

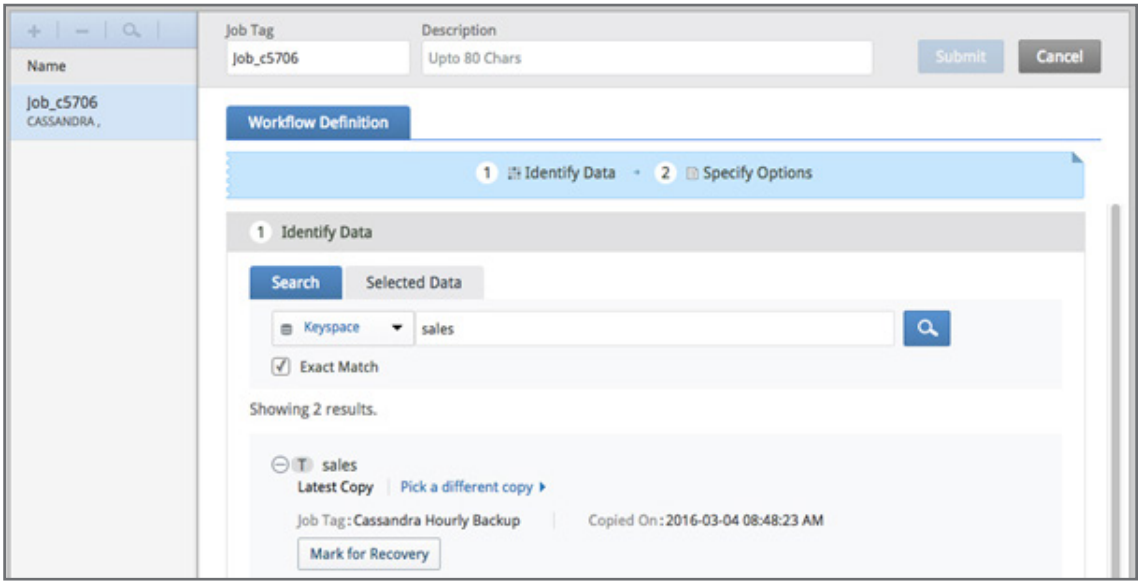
Backups performed using traditional solutions or Big Data vendor-supplied utilities are intolerant of topology alterations: in other words, any changes to the exact blend of servers and storage media that was present during the backup process can make a subsequent restore difficult - if not impossible - to carry out. Since the vast majority of Big Data environments are extremely fluid, this introduces a very tangible peril of permanent data loss because previous backups are likely to become useless.

Imanis Data follows a much more Big Data-friendly strategy. First, it dynamically acquires the topology of the destination cluster at the time of the restore. It then uses this structure to reshard the backed-up data to match the topology of its destination. This is a much smarter, more powerful mechanism for restoring information: it has no requirement for an exact match between the backup and restore topologies, nor does it require Imanis Data to preserve the production cluster's topology at the time of backup.

In addition, customers may elect to recover tables and keyspaces to the original source or to an alternate cluster. The destination cluster size is also independent and may be adjusted during the restore process.

Flexible data restoration

Imanis Data's recoveries are fast, granular, and capable of restoring data to any previous point in time no matter how large the backup data set may be. For example, although a Cassandra restore point may contain hundreds of tables and keyspaces, Imanis Data's FastFind technology lets users search for a specific keyspace and table, as well as mutations for that specific restore point. Similarly, FastFind can restore a single Hive partition out of the hundreds of partitions that this type of table may have.



Since Imanis Data uses an incremental-forever technique to instantly materialize backed-up data, recovery can take place very quickly: all that's necessary to perform a complete restore is the most recent checkpoint, since it already reflects all previous incremental backups.

The Imanis Data backup and restore architecture thrives when applied to the specialized platforms listed earlier. Rather than creating a "one size fits all" product, Imanis Data invested considerable time and resources to create a "data-aware" solution that not only backs up Big Data but also incorporates metadata such as file/directory attributes and table/ database schemas. This essential information is retained along with the restore point, and upon a user's request is then restored back to the production cluster with the data itself.

Table 1: itemizes the objects that Imanis Data assesses when performing a backup:

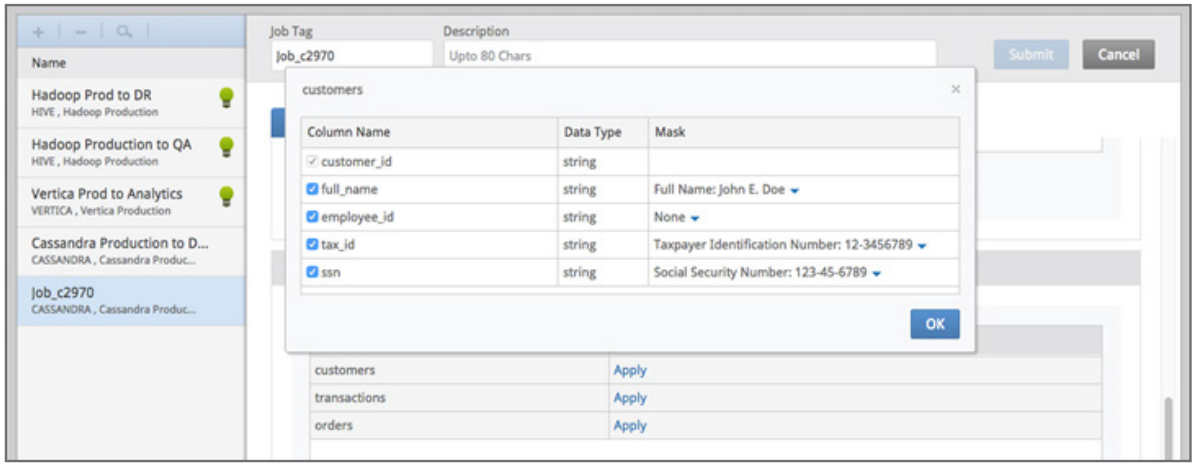
TECHNOLOGY	OBJECTS EVALUATED BY IMANIS DATA
Hadoop/HDFS	Files and directories – including permissions and attributes
Hive	Databases, tables, partitions, and metastore
Impala	Databases, tables, partitions, and metastore
Cassandra	Keyspaces and tables
Vertica	Databases, schemas, tables, and catalog
Couchbase	Buckets and documents
MongoDB	Buckets and documents

Making data available to software developers and testers

Protecting information is only part of Imanis Data's mission: as described earlier, software developers and testers need prompt and regular access to the production data that's stored in specialized Big Data platforms. This is a critical requirement to ensure that the new breed of applications that they're building will match user expectations. However, those enterprises that grant carte blanche access to this information confront a very serious risk of divulging Personally Identifiable Information (PII).9

To help eliminate this unpleasant possibility, Imanis Data uses its deep platform-specific knowledge to offer sophisticated data masking that lets organizations freely dispense Big Data yet protect PII. Imanis Data understands the data formats of each platform, so it's able to combine the details from the information source's schema and raw files to appropriately mask the relevant data. Currently supported masks include:

- Full name
- Social security number
- Taxpayer identification number
- Employee identification number
- Credit card number
- Email addesses



Imanis Data's data masks are always consistent. For example, a given social security number will always cause the same mask to be generated, and this coherence will occur across all tables containing columns with social security numbers. This preserves the statistical properties of the original data, and makes accurate analysis possible. Since making a full copy of production data for development and testing isn't always necessary or practical, Imanis Data lets users specify a meaningful sample size. It then uses this parameter to automatically drive the task of creating representative data in the target environment.

To further protect sensitive data, Imanis Data's patent-pending algorithm is completely one-way: there is no technique - even for Imanis Data - to reverse-engineer masked information. In addition, its data masking is stateless so there are no intermediate files or encryption technologies to maintain and protect.

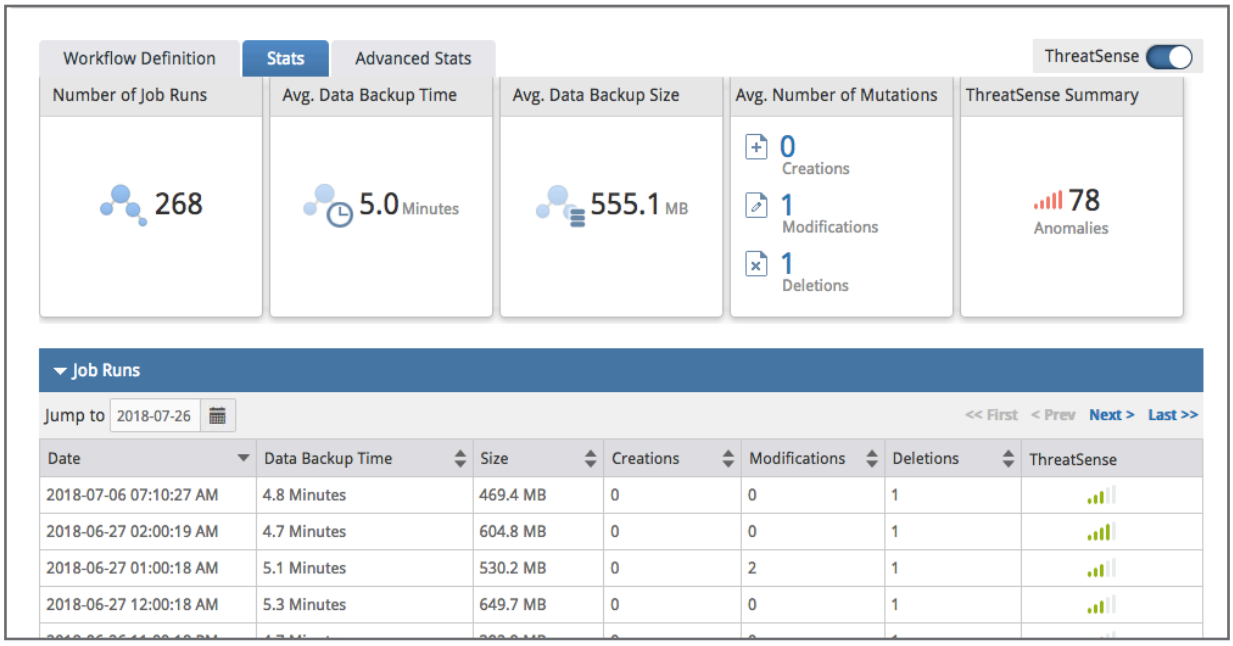
Finally, regardless of the exact data distribution configuration, Imanis Data offers self-service capabilities that let authorized users set up scheduled or ad-hoc production data extracts into development or testing sandboxes.

3. Machine Learning

When data volumes are small, decisions about when to back up data to meet business mandated RPO and RTO needs can reliably made by human judgement. Additionally, data anomalies such as data deletion can be easily detected since the environment is fairly simple. The margin of error resulting from human decisions is small and can be managed. However, when dealing with large volumes of data in complex distributed database environments, human decision becomes extremely difficult and error prone. Factoring in interactions between various systems is next to impossible in Big Data environments.

It is inevitable that best practices data management solutions not rely on human input and interaction to make decisions. Today, Imanis Data is already delivering machine learning ransomware detection with ThreatSense™, that:

- Provides ransomware and anomalous behavior detection
- Uses a novel unsupervised random forest mechanism for predictions
- Incorporates user feedback in subsequent learning



In the future, Imanis Data will expand their machine learning offerings in several areas:

- Optimizing Backup/Restore Performance – including determining best times to back up data, frequency of data backups, number of parallel backup/recovery streams, etc. in order to meet RPO and RTO needs.
- Cybersecurity – including detecting and recommending corrective action against data pattern anomalies created by ransomware attacks
- Optimizing Storage Usage – by determining data usage patterns and making decisions regarding data placement.

Using environmental data to answer these question is the future of data management. That is what Imanis Data is doing by leveraging machine learning and artificial intelligence to build sophisticated data models and driving all data management activity from the top using customers RPO, RTO, and compliance objectives.

SUMMARY & NEXT STEPS

As Big Data continues its march towards becoming a mainline IT resource, protecting this critical information and making it available for the builders of new applications will require fresh technologies and tactics. The Imanis Data platform manifests a set of intelligent design decisions that help enterprises deal with the three most well agreed-upon traits of Big Data environments:

Modern applications built upon Hadoop and NoSQL require a fresh approach to data management that addresses three key customer requirements:

Data Protection – of all Hadoop and NoSQL application data regardless of location to ensure organizations can recover from data loss and downtime caused by natural or man-made events.

Data Orchestration – flexible data mobility to copy, move, or migrate data to the appropriate locations on-premises or in the cloud based upon use-cases including test/dev (with masking and sampling), archiving, analytics, and compliance.

Automation – leverage machine learning to automate important data management tasks including detecting and recommending corrective action against ransomware attacks, optimizing backup/restore performance, and optimizing secondary storage efficiency by determining data usage patterns and making decisions regarding data placement.

About Imanis Data

Imanis Data is the machine learning data management company for the data driven world. The Imanis Data Management Platform enables customers to harness the full value of their data by delivering solutions that protect their data, as well as orchestrate and automate all their data management tasks. Imanis Data has been named a Gartner Cool Vendor as well as a CRN Emerging Vendor. Imanis Data's customers include leading Fortune 500 businesses in the retail, financial services and technology industries, among others. Backed by Canaan Partners, Intel, Onset Ventures, and Wipro Ventures Imanis Data is located in San Jose, CA.

Please contact us for more information at **info@imanisdata.com** or visit us at **www.imanisdata.com.**

IMANIS DATA

Imanis Data, Inc. | 2860 Zanker Road, Suite 109, San Jose CA 95134 | +1.408.649.6338 | imanisdata.com