Informatica

White Paper

Data Lake Management On Amazon Web Services (AWS) Reference Architecture

Building a Marketing Data Lake

ABOUT INFORMATICA

Digital transformation changes our expectations: better service, faster delivery, greater convenience, with less cost. Businesses must transform to stay relevant. The good news? Data holds the answers.

As the world's leader in enterprise cloud data management, we're prepared to help you intelligently lead—in any sector, category or niche. To provide you with the foresight to become more agile, realize new growth opportunities or even invent new things. With 100% focus on everything data, we offer the versatility you need to succeed.

We invite you to explore all that Informatica has to offer—and unleash the power of data to drive your next intelligent disruption. Not just once, but again and again.

Table of Contents

Introduction

Data is the lifeblood for any business, in any industry. However the ability to turn the right data at the right time into valuable business insights separates industry leaders from their competition. Data has become important in building efficient and effective marketing pipelines, as marketing technology transitions from traditional to digital marketing. In the new era of growing data demands, marketing analysts are driving the need for information technology departments to provide reliable, accurate, secure and governed data that will create new ways to interact with customers, build new data products or solutions, and generate new streams of revenue.

The Marketing Data Lake provides marketing teams access to information from multiple data sources in a single location where data was previously siloed into individual marketing platforms or unavailable due to the structure and size of the data. The Marketing Data Lake's self-service access enables marketing analysts to take a new approach to the integration, utilization, and analysis of their marketing data. By taking this approach to the collection, integration, and mastering of marketing-related data, marketing departments can establish micro-segmentation and personalization across multiple contact channels, initiate account-based marketing (ABM) practices to better understand the customer journey, and improve revenue generation across the organization.

Companies are looking for a framework to help implement a Marketing Data Lake that leverages new technologies such as Apache Hadoop either hosted on-premise or in the cloud for Big Data Analytics which can reduce infrastructure costs, increase customer loyalty, improve brand recognition and increase profitability. At the forefront of technology trends, Informatica's leading Data Lake Management solution allows organizations to confidently design and develop data lake architectures that deliver fit-for-purpose datasets for various analytics use cases.

Informatica's vision is to help organizations build a data-driven Marketing Data Lake platform by breaking down data silos, ensuring data analysts and data scientists autonomy in a fully secure and governed big data environment.

This reference architecture is designed to provide organizations a general framework that satisfies the implementation of a Marketing Data Lake using data lake management solutions and principles covering two current technology trends, Big Data and cloud computing. This document facilitates the design of a concrete Data Lake Management Reference Architecture by combining the effective tools, technology and approaches that will provide organizations flexible guidelines for establishing a solution architecture and solution implementation strategy.

Business Drivers

The business drivers for a Marketing Data Lake can vary between organizations and can be specific to an organization. However, a few generalizations are evident as trends across most organizations as key drivers responsible for a Marketing Data Lake.

Fragmented Marketing Data

Today, marketers have dozens of emerging marketing related SaaS application to work with. The analyst's primary challenge is working with fragmented data from siloed propriety SaaS application to successfully deliver customer experiences and drive sales revenue. Within an organization, separate teams often create data extracts or build specific data marts to facilitate building marketing analytics dashboards without sharing or cross-referencing datasets from the enterprise architecture. This dis-integration of data underscores the ability for a marketing team to analyze the data to customize products, personalize services or generate the best marketing leads. Therefore a central location is required to manage the supply and demand for this data, such as a Marketing Data Lake.

More Data, Big Data and Deeper Marketing Analytics

With marketing analysts having access to dozens of marketing related SaaS applications, each application has the potential to generate and consume increasing volume, velocity and variety of data, big data. The advent of big data analytics helps marketing analysts discover patterns, find trends and analyze data for overall marketing performance by evaluating key metrics for marketing effectiveness and return on investment.

Managing Big Data Governance and Quality

Data Governance and Quality are inherently a big data challenge. Much of the data that the organization uses, in the context of big data, comes from outside the organization with its origin or structure unknown increasing the likelihood of bad data. The moment that data is shared across business groups, it becomes an enterprise asset that must be governed and quality rules applied to maximize its value to the organization. Understanding the provenance of data through detailed data lineage instills confidence so that analysts can know whether to trust the data for their analytics. Customer related data, employee data, and data related to intellectual property of products and services must be identified and protected to comply with industry regulations.

Driving Principles of a Data Lake

In general, a data lake is a single Hadoop-based data repository to manage the supply and demand of data. Data Lake Management principles integrate, organize, administer, govern and secure large volumes of both **structured** and **unstructured** data to deliver actionable fit-for-purpose, reliable and secure information for business insights.

When embarking on a big data journey, it is important to recognize and keep in mind a few principles required of a data lake.

- There are no barriers to onboard data of any type and size from anywhere
- Data must be easily refined and immediately provisioned for consumption
- Data must be easy to find, retrieve, and share within the enterprise
- Data is a corporate accountable asset, managed collaboratively by data governance, data quality and data security initiatives

Key Characteristic of Data Lake Management

Marketing data is generated from many various sources: website activity, lead generation data, customer relationship management (CRM) applications, transaction, social media, geo-location and third party data. The diversity of the data types comes in a variety of formats: unstructured data (e.g. Twitter, LinkedIn, Facebook), semi-structured data (e.g. Salesforce, Marketo, Adobe Analytics) and structured data (e.g. Marketing Database Applications). These marketing data types must be integrated and stored in their native format which facilitate concepts like schema-on-read on low cost commodity hardware that is hosted on-premise or in the cloud. Big data processing frameworks allow organizations to process massive amounts of marketing data at scale, benefitting from parallelization of data processing tasks using distributed algorithms on big data storage/processing platforms such as Hadoop.

Data Lake Management is the process of integrating, cataloging, cleansing, curating, governing and securing data assets on a big data platform. The goal of data lake management is to ensure data is successfully managed from ingestion through consumption. A Marketing Data Lake initiative requires a disciplined approach to data lake management practices, tools, platform and policies to architect, develop, test and monitor key components.

As a multi-tiered data platform, the Marketing Data Lake enhances the data scientists' ability to perform successful marketing analytics and operational activities leveraging self-service Intelligent Data Applications that facilitate agile analytics and improve actions leading to successful outcomes.

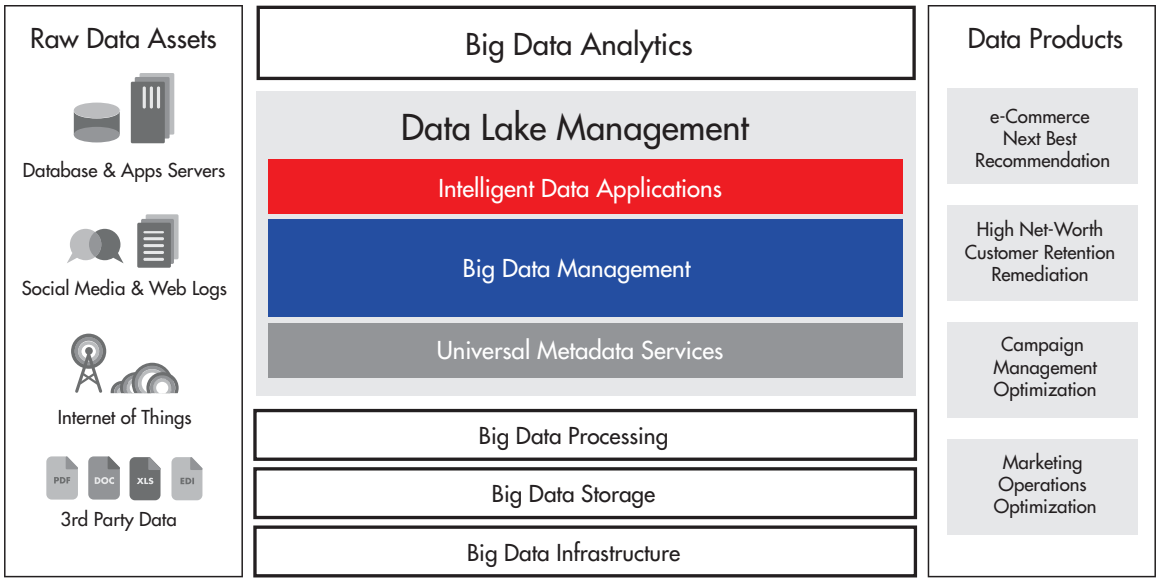

Figure 1. Data Lake Management Core Components.

A Marketing Data Lake reference architecture must include core components of a data lake management platform identified as Intelligent Data Applications, Big Data Management and Universal Metadata Services.

Intelligent Data Applications provide data analysts, data scientists, data stewards and data architects with a collaborative self-service platform for data governance and security that can discover, catalog and prepare data for big data analytics.

Big Data Management is the process of integrating, governing and securing data assets on a big data platform and has three sub pillars, defined as:

- **Big Data Integration** is the collection of data from various disparate data sources, such as Salesforce, Marketo, Adobe, and an Enterprise Data Warehouse, which is ingested, transformed, parsed and stored on a Hadoop cluster to provide a unified view of the data. Big Data Integration plays an important role as it allows data engineers to extract data from various marketing applications, apply business logic as defined by data analysts and load the data to a big data store such as Hadoop.
- **Big Data Governance** relates to managing and controlling data while Quality is the ability to provide cleansed and trusted data that can be consumed or analyzed by an intelligent data application or a big data analytics tools. Data stewards must manage and maintain the ever increasing variety of data, specifically customer data, by breaking down the taxonomy of each marketing data type at a granular level and curating the data into a reusable, consumable, data asset.

- **Big Data Security** is concerned with protecting sensitive data across the Marketing Data Lake. Security analyst teams need to discover, identify and ensure any customer data stored in weblogs, CRM applications, internal databases, third-party applications are protected based on defined data security policies and practices. The team needs full control and visibility into any data in the Data Lake and the ability to monitor for unusual behavior or non-compliance. Additionally, sensitive customer data must be masked to prevent unauthorized users from accessing sensitive information.

Universal Metadata Services manages all the metadata from a variety of data sources. For example, an enterprise information catalog manages data generated by big data and traditional sources by collecting, indexing, and applying machine learning to metadata to provide metadata services such as semantic search, automated data domain discovery and tagging, and data intelligence that can guide user behavior.

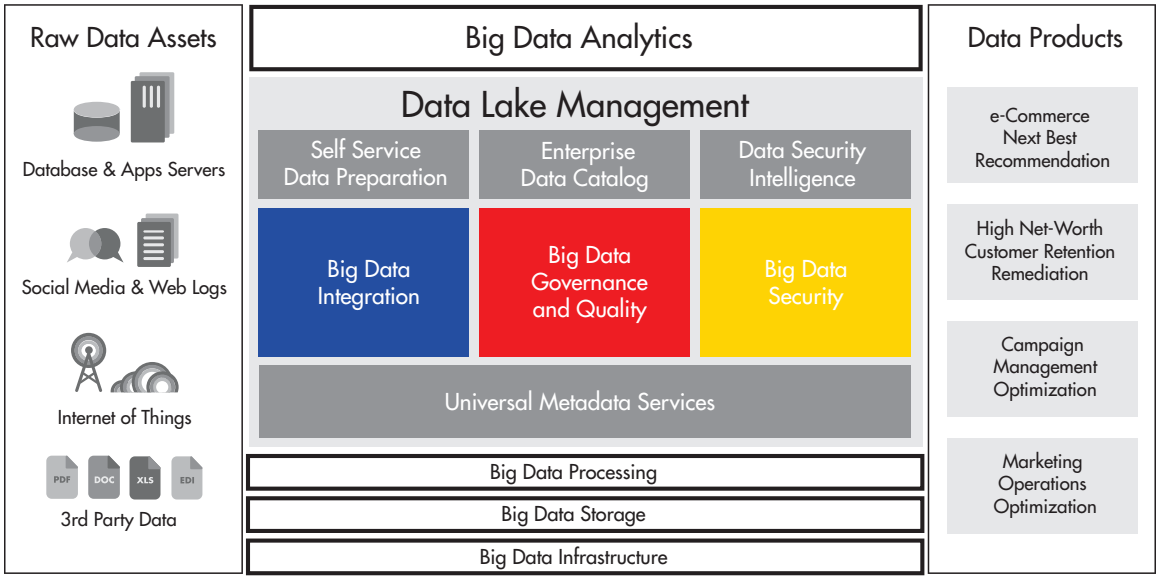

Figure 2. Intelligent Data Applications & Big Data Management key sub-components.

Key Capabilities of Data Lake Management

Within the three core components of data lake management, figure 3 highlights capabilities used to define a big data architecture for a Marketing Data Lake solution. Each capability is described in the following sections.

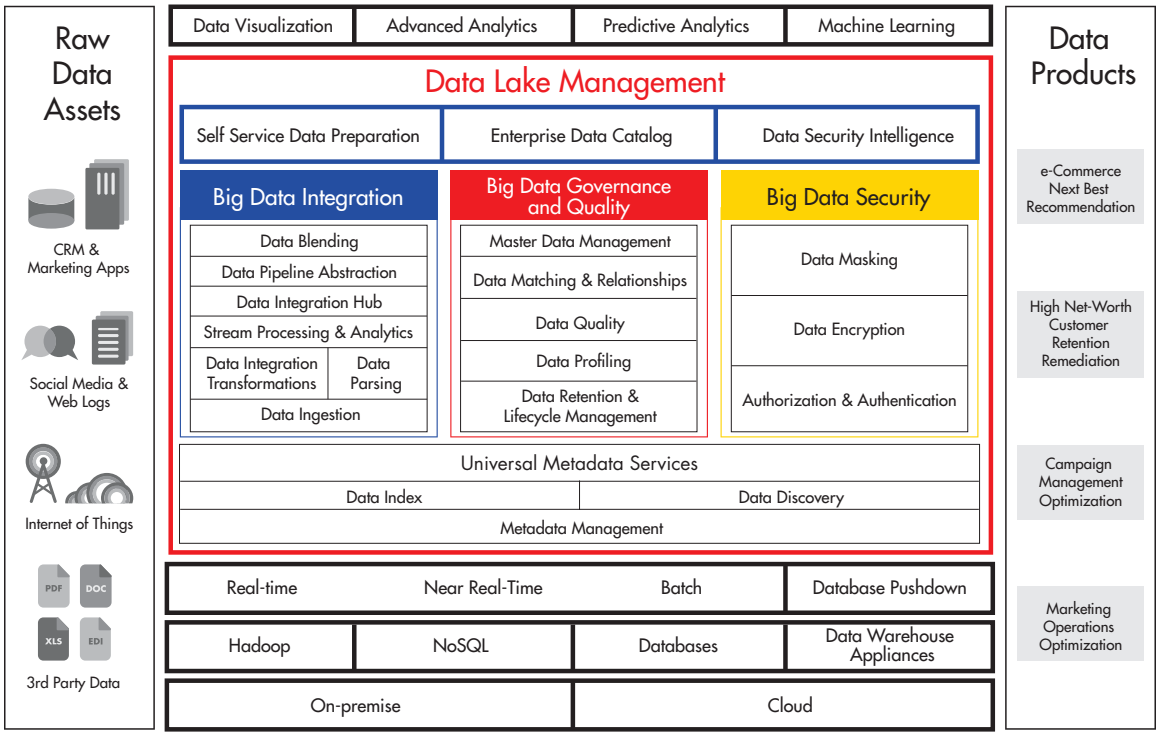

Figure 3. Key Data Lake Management Capabilities.

Raw Data Assets – Data collected from source systems without any preprocessing or data manipulations. Raw data may include data from data warehouses, CRM application such as Salesforce, web analytics data from Adobe Analytics or Google Analytics.

Big Data Infrastructure – From a data center perspective (e.g. on-premise, cloud) accommodates the increasing volume and variety of data types reliably and having the ability to scale-up (increasing individual hardware capacity) or scale-out (increasing infrastructure capacity linearly for parallel processing), either deployed on-premise, in the cloud or as a hybrid solution.

Big Data Storage – Provides the ability to store large amounts of a variety of data (structured, unstructured, semi-structured) at scale delivering the necessary performance as measured by Input/Output Operations Per Second (IOPS), to guarantee timely delivery of data. Typically, these environments will run Hadoop or NoSQL which provide scalable and reliable distributed data storage spanning large clusters of commodity servers. Traditional data storage & processing platforms such as database or data warehouse appliances (Teradata, Netezza, Greenplum) complement Hadoop & NoSQL.

Big Data Processing – Provides the ability to process data at any latency using big data processing frameworks. There are three categories related to data process latency: real-time, near real-time and batch.

- **Real-time processing** refers to the continuous processing of data. Typical processing frameworks include Apache Kafka.
- **Near Real-time processing** refers to data processed at low latency set intervals rather than instantaneously using Spark Streaming or Apache Storm.
- **Batch processing** refers to data that is typically processed at latencies on the order of seconds, minutes or sometimes hours using Apache MapReduce, Apache Tez, or Apache Spark.

The Marketing Data Lake requires an implementation of each latency type to solve different use cases. For example, organizations can manage and monitor customer interactions or customer feedback in real time while customer addresses are updated once on a nightly basis.

Universal Metadata Services – A key capability for a big data initiative is the ability to catalog all data, enterprise-wide regardless of form (variety) or where it's stored, on Hadoop, NoSQL or an Enterprise Data Warehouse, along with the associated business, technical, and operational metadata. The catalog must enable business analysts, data architects, and data stewards to easily search and discover data assets, data patterns, data domains, data lineage and understand the relationships between data assets – a 360 degree view of the data. A catalog provides advanced discovery capabilities, detailed data lineage, smart tagging, data set recommendations, metadata versioning, a comprehensive business glossary, and drill down to finer grained metadata.

- **Metadata Management** – Effectively governed metadata provides a view into the flow of data, the ability to perform impact analysis, a common business vocabulary and accountability for its terms and definitions, and finally an audit trail for compliance. The management of metadata becomes an important capability to oversee changes while delivering trusted, secure data.
- **Data Index** – Enables data analysts or data scientists to find and understand the relationship of the data. The data index empowers data analysts to easily collaborate by tagging, annotating and sharing data assets and contribute their knowledge about the data back to the data index.
- **Data Discovery** – Allows a marketing team to find data patterns, trends, relationships between data sets (data domains, join keys) and entities (e.g. householding), and anomalies or business scenario occurrences across all data sources. For example, the data security team may need to identify where Personally Identifiable Information (PII) is used and how that relates to specific business rules and processes where data masking can then be used to protect customer data.

Big Data Integration – The Data Lake architecture must integrate data from various disparate data sources, at any latency, with the ability to rapidly develop Extract Load Transform (ELT) or Extract Transform Load (ETL) data flows and deploy anywhere on Hadoop, in a data warehouse, on-premise or in the cloud. Key capabilities for Big Data Integration are noted below:

- **Data Ingestion** – Ingest data from any source (marketing databases, Marketing SaaS Application, weblogs), at any speed (real-time, near real-time or batch) using high-performance connectivity through native APIs to source and target systems with parallel processing to ensure high-speed data ingestion and extraction.
- **Data Integration Transformation** – Provides data engineers access to an extensive library of prebuilt data integration transformation capabilities that run on Hadoop.
- **Data Parsing** – The ability to access and parse complex, multi-structured, unstructured data such as weblogs, JSON, XML, and machine device data.
- **Data Integration Hub** – A centralized publish/subscribe model hub based architecture for agile and managed enterprise data integration.
- **Data Pipeline Abstraction** – The ability to abstract data flows from the underlying processing technology, such as MapReduce, Spark, or Spark Streaming, insulating the implementation from rapidly changing big data processing frameworks and to promote re-use of data flow artifacts.
- **Data Blending** – Provides data analyst or data scientist self-service capabilities to rapidly ingest data, discover patterns, merge data, transform trusted data to gain deeper insight into marketing activities or develop other marketing related data sets.

Big Data Governance & Quality is critical to the Data Lake especially when dealing with a variety of data. The purpose of big data governance is to deliver trusted, timely, and relevant information to support the marketing analytics. Big Data Governance provides the following capabilities:

- **Data Profiling** helps data analysts understand the structure, completeness, and relevance of data to ensure consistency enterprise-wide while identifying and remediating bad data.
- **Data Quality** is the ability to cleanse, standardize, and enrich data in the Marketing Data Lake using pre-built data quality rules and techniques.
- **Data Matching & Relationships** is the ability to match and link entities such as customers or products within and across multiple sources and link them together to create a single view of a data. Sophisticated algorithms identify insightful relationships such as for householding.
- **Master Data Management** is the creation of single authoritative master records for all critical business data (e.g. customers, products, suppliers), leading to fewer errors and less redundancy in business processes.
- **Data Retention & Lifecycle Management** is the ability to automatically archive redundant and legacy application data for regulatory compliance and to control costs while maintaining necessary access to inactive data.

Big Data Security is the process of minimizing data risk including discovering, identifying, classifying, and protecting sensitive data, as well as analyzing its risk based on value, location, protection, and proliferation. The following capabilities define big data security:

- **Authentication** is used to determine whether a user has access to the Marketing Data Lake, while authorization determines if a user has permission to a resource within the Marketing Data Lake.
- **Data Encryption** is the ability to encrypt data using encryption algorithms using an encryption key making it unreadable to any user that doesn't have a decryption key.
- **Data Masking** is the ability to de-identify and de-sensitize sensitive data when used for support, analytics, testing, or outsourcing.

Intelligent Data Applications continuously discover, learn, and understand enterprise data as it exists in the data lake and the enterprise data warehouse determining relationships, quality and usage patterns from every interaction to measure risk and trust. Examples of unique intelligent application capabilities are defined as follows:

- **Self-Service Data Preparation** provides analysts an easy to use self-service user-experience with collaborative data governance which guides behavior through intelligent recommendations to quickly prepare data for analytics.
- **Data Catalog** provides a rich set of capabilities for enterprises to catalog, manage, curate, and maximize the value of their data assets by automatically analyzing and understanding large volumes of metadata within the enterprise.
- **Data Security Intelligence** enables security analysts to discover, classify, and score sensitive data across the enterprise based on security policies.

Big Data Analytics provides capabilities for data analysts and data scientists to apply advanced analytics using trusted, governed and secure data to discover valuable insights, improve decision-making, and minimize risks using various data visualization, advanced statistics, predictive analytics, and machine learning tools.

Data Products operationalize analytic models using automated data pipelines that transform raw data and feed end-user applications with actionable insights.

Informatica's Technology Supporting the Marketing Data Lake

Informatica's Data Lake Management Solution brings together the core capabilities (described in Figure 1) intelligent data application, big data management and universal metadata services into a single integrated platform. The Marketing Data Lake provides an integrated, end-to-end view across all the marketing related data.

The high-level diagram below (Figure 4) illustrates how Informatica's Data Lake Management solution and components support the marketing data lake. The diagram is color coded to match Informatica products to the key data lake management capabilities.

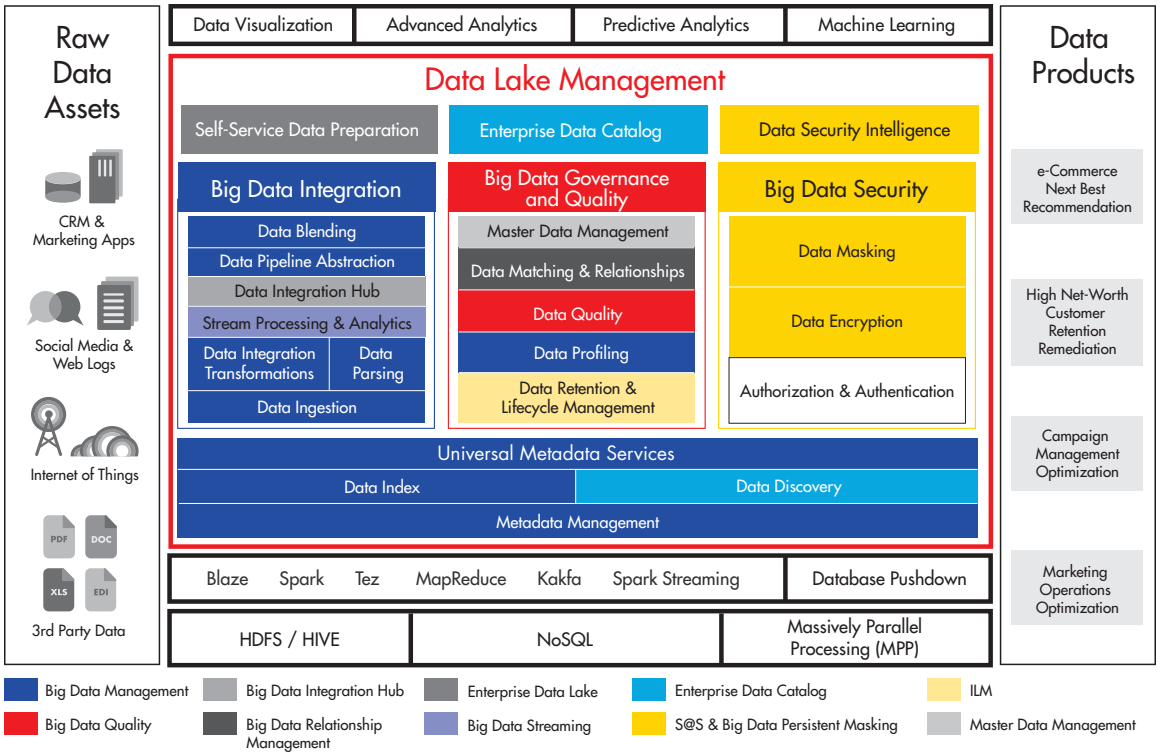

Figure 4. Informatica's Data Lake Management Solution support for the Marketing Data Lake.

Described below are Informatica Data Lake Management Solution used to support the Marketing Data Lake:

Big Data Management (BDM)

BDM delivers a comprehensive big data integration solution to natively ingest, transform and integrate big data workloads in Hadoop.

Big Data Streaming

Informatica Big Data Streaming allows organizations to prepare and process data in streams by collecting, transforming and joining data from a variety for sources scaling for billions of events with a processing latency of less than a second. Big Data Streaming leverages prebuilt transforms which run natively on Spark Streaming and Apache Kafka as the data transport across mappings and data replay for recoverability.

Big Data Quality (BDQ)

BDQ delivers a comprehensive big data solution to natively profile, discover, and cleanse big data workloads in Hadoop.

Big Data Relationship Manager (BDRM)

BDRM matches and links any type of data on Hadoop to develop a 360 degree view of data. BDRM uses advanced algorithms to match entities and to discover relationships such as for householding.

Big Data Integration Hub (BDIH)

BDIH is a centralized modern hub based application that automates the publish/subscribe data pipelines between SaaS, Cloud and on-premises applications and the marketing data lake.

Enterprise Data Lake

Enterprise Data Lake provides self-service efficiency for business analysts and data scientists to prepare and collaborate on data for analytics by incorporating semantic search, data discovery and intuitive data preparation tools for interactive analysis with trusted, secure and governed data assets.

Enterprise Data Catalog (EDC)

EDC enables business analysts, data architects, and data stewards to easily search and discover data properties, data set patterns, data domains, data lineage and 360 data relationships. It provides advanced discovery capabilities, smart tagging, data set recommendations, data similarity, metadata versioning, a comprehensive business glossary, and drill down to finer grained metadata.

Secure @ Source (S@S)

S@S enables security analysts to discover, classify, and score sensitive data across the enterprise. Data security intelligence dashboards monitor sensitive data protection, proliferation, and usage along with alerts to proactively notify security analysts of suspicious behavior or of specific data security risks based on policies which provide a more risk-centric approach to protecting sensitive data.

Big Data Persistent Masking

Big Data Persistent Masking helps IT organizations manage the access to their most sensitive data. Data masking protects confidential data such as credit card information, social security numbers, names, addresses, and phone numbers from unintended exposure to reduce the risk of data breaches. It provides unparalleled enterprise-wide scalability, robustness, and connectivity to a vast array of databases.

Master Data Management (MDM)

MDM achieves and maintains a consistent view of master data across an enterprise, including the relationships across the master data (e.g. customer householding). MDM consolidates (either through linking records or merging them), resolves conflicting data sources and establishes a trusted, authoritative source of reference for commonly-used master data.

Information Lifecycle Management (ILM)

ILM provides a comprehensive approach to managing structured data in databases, enterprise applications, data warehouses and files—including reference, transactional and related unstructured data such as images, audio, documents and other types of attachments. The ILM solutions help organizations manage data growth in both production and non-production environments.

Data Lake Management Reference Architecture for the Marketing Data Lake

The reference architecture takes a holistic approach to building a marketing data lake that reflects a well-defined environment incorporating the data lake management framework components described in the previous sections. The Marketing Data Lake architecture comprises of five distinct functional areas: landing zone, streaming analytics, data curation process, discovery zone and enterprise zone and four foundation areas: Metadata Management, Data Governance, Data Security and Data Infrastructure. Segmenting the data lake into zones allows storage and organization of all the data from across the enterprise while augmenting existing systems, such as the enterprise data warehouse or Master Data Management Hub. This provides a consolidated view of data with relevant facts allowing the data analyst or data science team to reliably and consistently find, share or curate all types of data.

The next sections discuss the Data Lake functional areas shown in Figure 4 and how they relate to the key capabilities of a data lake management framework in support of the Marketing Data Lake architecture, the data process flow within each zone and the roles of IT and business.

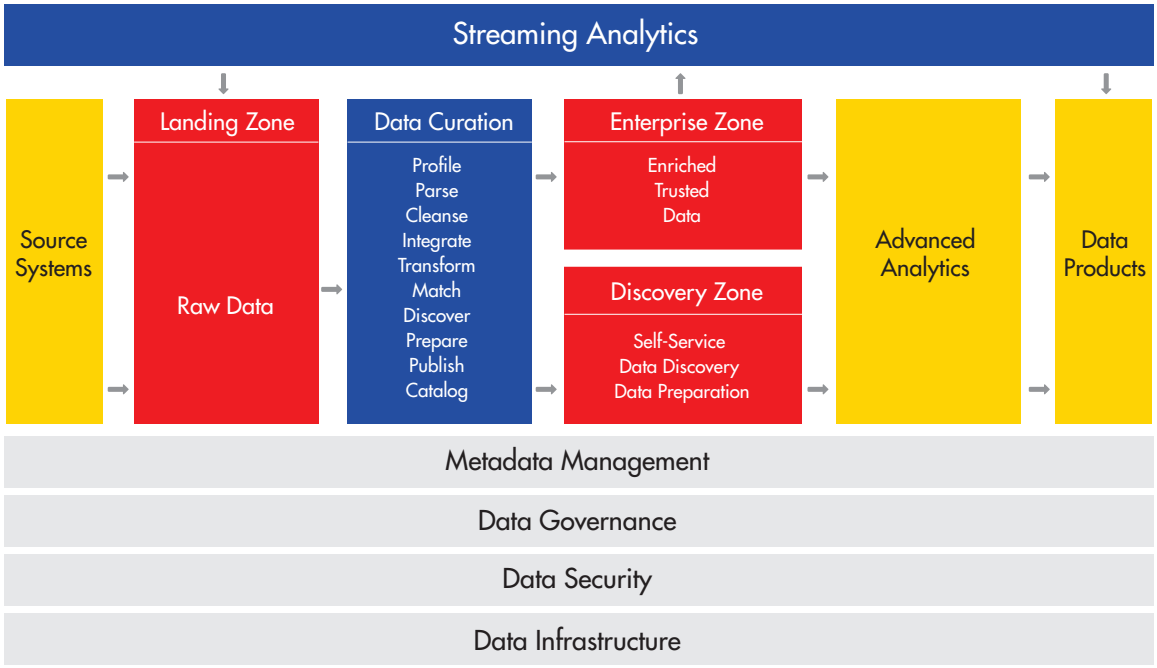

Figure 5. The Informatica Marketing Data Lake Reference Architecture.

The Landing Zone

The Landing Zone is dedicated to raw data storage that will serve as a repository of historical data for the organization. The landing zone typically contains structured, semi-structured as well as unstructured data. Structured data, typically from a Data Warehouse, is offloaded to Hadoop which involves offloading ELT processing and infrequently used data to Hadoop. This alleviates the CPU burden on data warehouses consumed by in-warehouse data transformations in ELT models, and frees space by offloading low-value or infrequently used information. This use case is also known as Data Warehouse Optimization for Hadoop.

The landing zone often receives data from a variety of sources such as relational, social media, SaaS applications, mainframe, and clickstream and weblogs. Data Ingestion is the primary activity that occurs in the landing zone where data is stored in its native format of the data source.

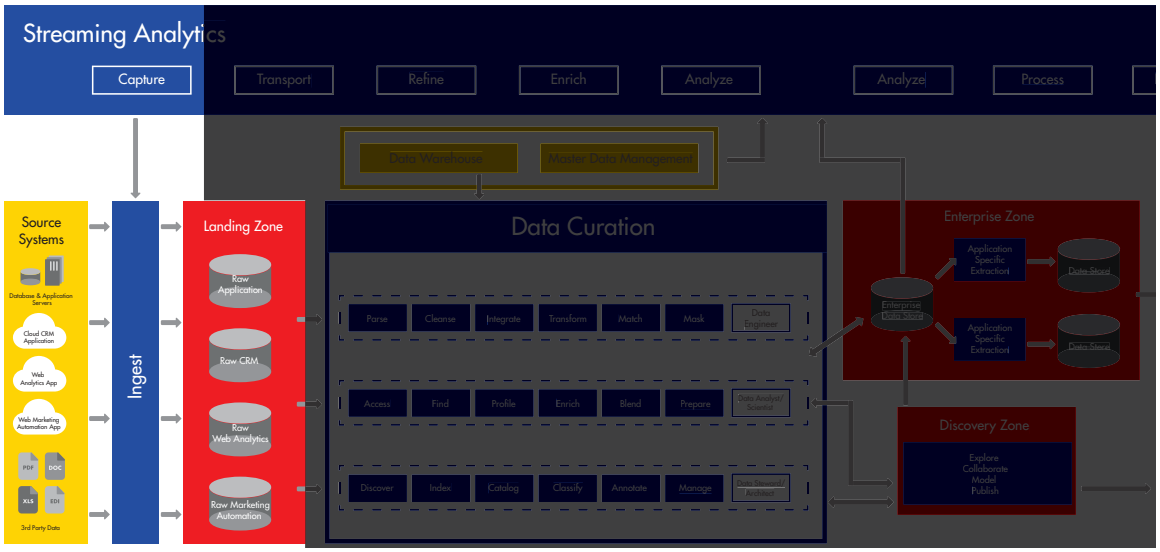

Figure 6. The Landing Zone within the data lake.

Big Data sources like social media, clickstream, and sensor data from IoT have no system of record as the data is typically generated outside of the organization—the Landing Zone becomes the system of record for these data types. Other more traditional data sources, like an organization's transactional and operational data, have their own system of records, but the applications that generate this data do not always persist the history. The Landing Zone can become the historical record for this type of data (sometimes referred to as an active archive).

It is imperative that data is loaded into the landing zone in its raw, untransformed format so that a historical record of what the data represented at the time it was captured can be maintained. As data is stored in its raw native format in the Landing Zone, data quality rules to standardize and validate the data are not yet applied.

Data Engineers must build scalable, reusable data flows to ingest a variety of data into the landing zone. Data engineering teams are responsible for the design, build, and deployment of data flows, and validating the results within this zone. IT is also tasked with the responsibility for managing data security and the data infrastructure. Data analysts or data scientists can access raw data from the landing zone or upload data to the zone as they work within the discovery zone (detailed below).

Access to the landing zone is typically highly restricted because data in its raw form is difficult to interpret and sensitive information might be exposed in its raw form.

The Data Curation Process

The next process in the marketing data lake is dedicated to the data curation lifecycle which consists of three phases: management of all data elements and data lake architectures, self-service data preparation and big data integration. Additionally, data from the Enterprise Data Warehouse or Master Data Management Hub can be augmented to the data curation process to provide a consolidated view of the data.

The data curation process enables collaboration between data engineers, data analyst and data stewards to build a holistic view of the organizations data within the marketing data lake. Described below is the big data curation lifecycle and the roles responsible for each process.

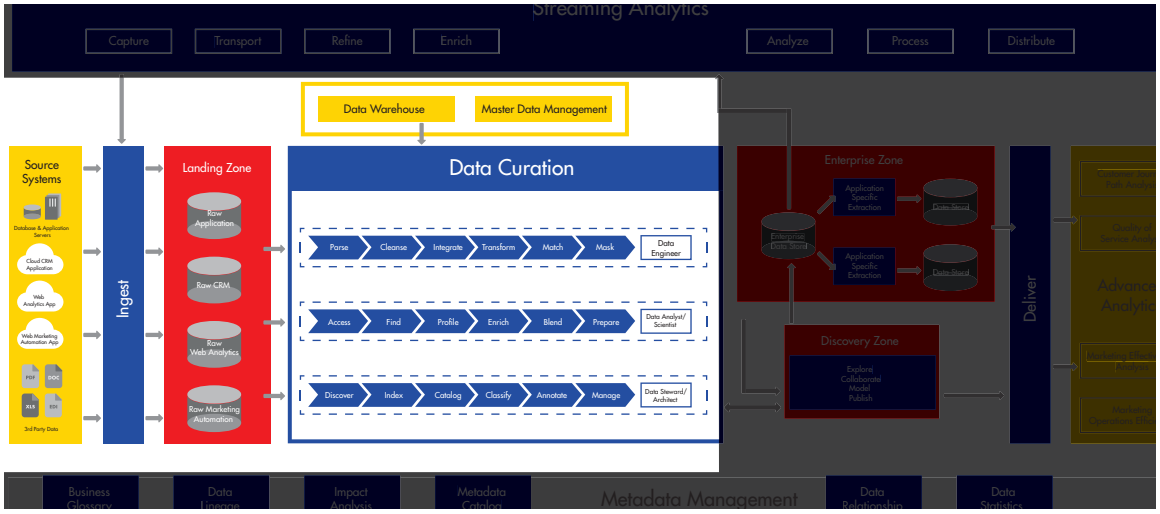

Figure 7. The Data Curation Process in the marketing data lake.

The first order of building a successful marketing data lake is to understand the data and metadata of data assets ingested in the landing zone and within the enterprise.

Data Stewards are responsible for managing data assets in the data lake and the enterprise ensuring high levels of data quality, integrity, availability, trustworthiness, and data security while emphasizing the business value of data. By building a catalog, classifying metadata and data definitions, maintaining technical and business rules and monitoring data quality, data stewards ensure data in the lake is consistent for use in the discovery zone and enterprise zone. As the inventory of technical and business metadata is collected and indexed and data sets become available, data architects can design robust scalable data models and flows to meet the business goals of the marketing data lake.

The data lake's self-service access allows data analysts to find, profile, enrich, blend, prepare and share data assets that reside in the marketing data lake or outside the data lake to deliver sustainable business value. The following describes self-service data preparation steps undertaken by a data analyst team:

- Access & Find – Through a defined metadata catalog, a data analyst can easily find, access and explore any data assets in the enterprise or in the data lake.
- Profile – A data analyst reviews the structure, completeness, accuracy, duplicates, integrity and relevance of data to ensure consistency enterprise-wide while identifying and remediating bad data. Analysts can build and apply business rules to identify and resolve any data issues.

- Enrich – After identifying data quality issues thru data profiling, data analysts can match, standardize or enhance raw data. For example, augmenting web analytics application data with geographic or demographic data, allows marketing analysts to target different segments of the population by area and preferences.
- Blend & Prepare – Data Analysts can prepare data by combining, merging, transforming, cleansing and structuring data from one or more data assets making it available for analysis for the discovery zone or they may publish the curated data to the enterprise zone.

Using the defined data lake architecture, the data engineering team is responsible for the design, build, and deployment of the project's big data integration phase. The team is responsible for ensuring data ingested into the landing zone from various source systems adheres to the original data formats, data quality standards defined by data stewards and complex data transformations are applied to data sets within the marketing data lake for use by data scientists during exploratory analysis. The following describes the big data integration phase when designing complex data flows:

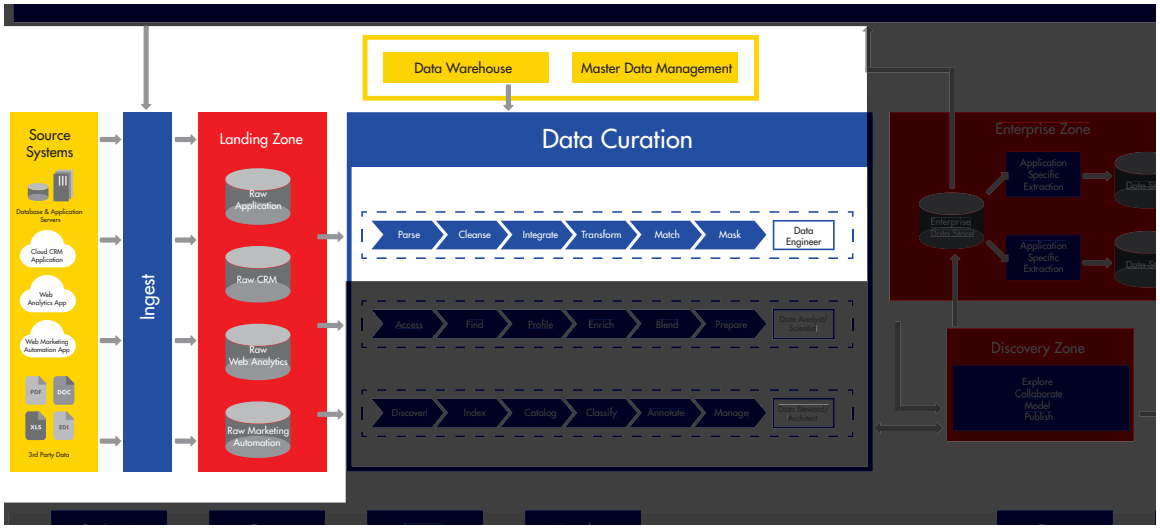

Figure 8. The Big Data Integration phase in the marketing data lake.

- Parse – Big data generated from social media, weblogs, clickstream or sensor and device data must be converted from multi-structured or unstructured data to a usable & readable format for analysis. Data engineers use data transformation tools to parse, map, serialize and split data for unstructured, semi-structured and complex Hadoop formats such as Avro, Parquet or ORC. For example, parsing a weblog from unstructured data and extracting features into a Hadoop format such as Parquet to optimize the performance of the queries or optimize to load data fast to the data lake.
- Cleanse – Once the data analyst or data steward identifies issues with data during profiling, data is cleansed, standardized and conformed to the level required for the analytics. Examples include address cleansing or validation, date and code standardization.
- Integrate & Transform – The objective of the integrate & transform phase is to join cleansed data sets and apply complex transformation rules to data sets for usage in the enterprise data zone. Additionally, Data Quality rules (identified in the profile phase) or data masking rules can be integrated into a data pipeline. Existing data from the enterprise data warehouse, master data management hub or the enterprise zone can be augmented in this phase adding more value to the data in the data lake.

- Match – As data is integrated and transformed, duplicate records can be identified using a variety of sophisticated matching techniques. Matching can occur within a single data set (e.g. identify duplicate customer records) or match entities (e.g. customers) between two datasets when merging or appending new data.
- Mask – Depending on the use case, data can be masked to prevent unauthorized users from accessing or viewing sensitive information. For example, masking Personally Identifiable Information (PII) customer data such as social security number or biometric information before the data is consumed or used by an analyst or downstream application.

The data refined during the data curation process is stored in the Enterprise Zone and also made available to the data discovery zone.

The Discovery Zone

The key to unlocking value from a marketing data lake is to give data analysts and data scientists the ability to quickly and iteratively analyze the raw data, augmented with enterprise data, so they can look for "gold nuggets" that deliver new insights.

The discovery zone is designed as a self-service layer for exploratory analysis and rapid iteration that allows data analysts to collaborate and explore data and quickly view data lineage, view relationships between data assets, and prepare and publish data for consumption by the advanced analytic zone or enterprise zone.

The Enterprise Zone

The enterprise zone is the centralized location in the data lake for enriched and trusted data. The enterprise zone is the single point of reference within the data lake (discovery, data curation process & advance analytics) and can be shared across the enterprise for big data analytics.

High-value data generated from this zone can be used to update data in the Enterprise Data Warehouse or used during the data curation phase or consumed by analytic applications.

Streaming Analytics

Streaming analytics relates, for example, to streaming clickstream or weblog or device data for real-time analytics. Streaming data can be persisted in the landing zone for historical batch analysis. Depending on the velocity of the data stream and size of data packets a data engineer may choose to sample the data at a lower rate to reduce the amount of data stored in the data lake. Within this phase, streaming data is transformed (filtered, aggregated, and enriched) and can be enriched with other data from the enterprise zone in real-time, sending real-time alerts or notifications for immediate action. Data engineers can deploy trained models built by data scientists during the exploratory phase to the streaming analytics layer.

Advanced Analytic Zone

Using advanced analytical techniques, data scientists can model or train data sourced from the landing zone or enterprise zone to analyze data to make predictions, make strategic decisions based not only on what has occurred but on what is likely to occur, and to generate recommendations turning valuable insights into business-improving actions.

Metadata Management

Metadata Management, one of the core foundations of the data lake, allows an organization to manage the entire data lifecycle providing data transparency and visibility. Data Lineage visualization and profiling statistics help data stewards and data architects visualize the provenance and quality of data in the data lake and enterprise and perform impact analysis due to changes in data definitions and rules.

Data Governance

As discussed earlier, Data Governance is critical to the Marketing Data Lake and the purpose of big data governance is to deliver trusted, timely, and relevant information to support marketing analytics. Data Governance is the discipline that creates repeatable and scalable data management policies, processes and standards for effective use of data. Data governance requires business participation so that appropriate controls, processes and methods can be developed for data stewards and other data custodians right from the start.

Data Security

Understanding sensitive data risk is key to protecting business critical data; the process for analyzing data risk includes detecting, identifying, and classifying it, as well as analyzing its risk based on value, location, protection, and proliferation. Once the type and level of risk have been identified, data stewards and security analysts can take appropriate steps such as data masking to ensure data is safe and secure.

Data Infrastructure

As described in the key capabilities section, the data infrastructure relates to the big data storage (such as Hadoop & NoSQL) and processing frameworks (MapReduce, Spark, Spark Streaming). The big data infrastructure is deployed and maintained with the help of leading Hadoop vendors (Cloudera, Hortonworks, MapR, Amazon AWS, Microsoft Azure, IBM Big Insights) which can be deployed on-premise or in the cloud.

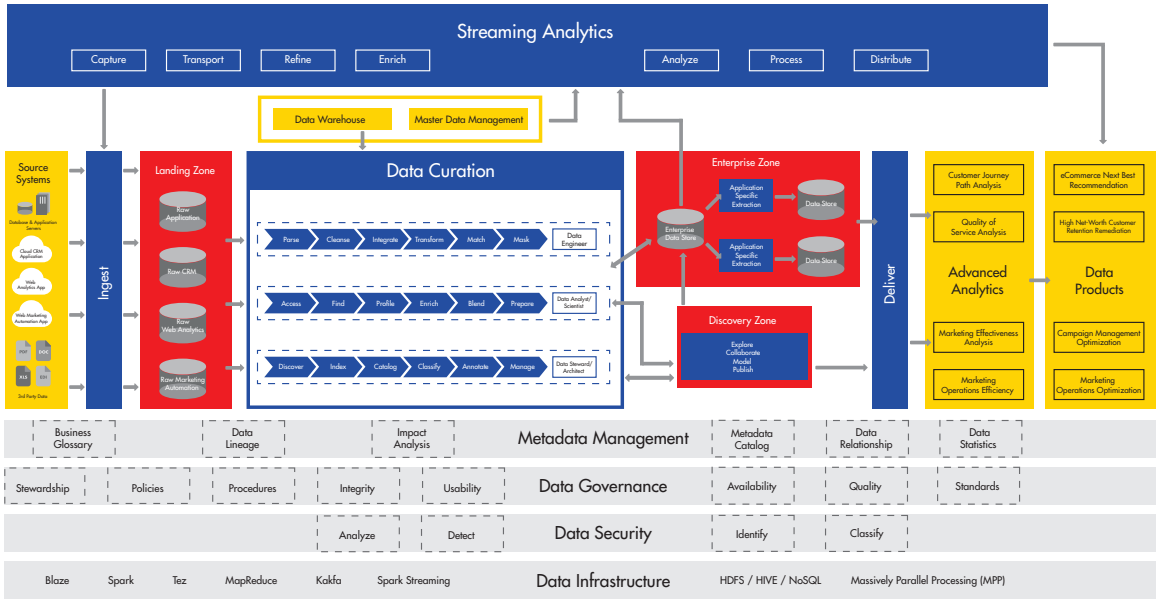

Figure 9. The Enterprise Zone within the data lake. This shows the end state of the Data Lake Management Reference Architecture.

Becoming Big Data Ready

Implementing a data lake without a strategic plan that includes the key components of data lake management and the right talent, big data projects can easily become a liability rather than a valuable asset. Commitment from the business sponsor-executives is critical for the success of big data projects. Big Data projects are a collaborative effort with several team members that include business subject matter experts, data analysts, data scientists, data stewards, data engineers, and data architects. The business understands the data, while IT knows how to implement processes to store and process the data in the data lake. Architects must consider the following best practices for building a data lake using the data lake management framework.

- **Centralize the people, standardize the process, and drive consistency in the architecture.** A centralized data management team enables efficiency and agility for analytical users. A standardized process and consistent architecture ensures that your organization's innovation is fully focused on compelling analytics.
- **Drive collaboration between data management and data analyst teams with small core project teams that iterate on requirements and deliver with agility.** Waterfall requirements gathering can impose too many hurdles for big data, and bottleneck the fast pace of business. Cross-functional project teams with shared priorities can enable agility.
- **Establish taxonomies and classifications to easily develop fit-for-purpose data available to analysts.** Sufficient preparation of raw data into consistently parsed and structured data dramatically reduces the overhead of data preparation by analytical users. Standardized taxonomies can also simplify finding, auditing and lineage tracking for compliance and governance.
- **Use a data lake to curate fit-for-purpose data and collaboratively drive data quality.** Big Data analytics teams can't wait for lengthy software development life cycles (SDLC) associated with traditional data warehousing before getting access to data. A well-managed data lake quickly provisions minimally transformed data that is fit-for-purpose. Analytical users can then collaboratively filter, join, aggregate, prepare, and curate data increasing the value of the data in the lake.
- **Assess data quality and deliver proactive data quality visibility.** Instrument the data lake to automatically detect signals of incomplete and inconsistent data so that data quality rules can be applied as needed. Use data quality scorecards and dashboards to drive visibility and understanding of data assets in the lake.
- **Use curated data for secure and governed access by other systems.** After data is aggregated, compressed, partitioned, matched, and masked, it can be used as a trusted and tracked source for high quality reporting by a wider variety of users.

Deployment Reference Architecture for Amazon AWS

This section provides guidance on how to deploy Informatica Big Data Management (BDM) on Amazon Elastic Map Reduce (EMR) to support a Marketing Data Lake.

This section assumes you have a conceptual understanding and some experience with Amazon EMR, Hadoop and Informatica Big Data Management.

Organizations are looking for opportunities to reduce their on-premise datacenter footprint by offloading or extending on-premise applications, data warehouses, big data solutions to the cloud. Cloud deployments increase agility as they allow organizations to rapidly add new capabilities and scale up and down as their needs change. Cloud solutions free up IT resources from supporting commoditized infrastructure and allow them to focus on building differentiated capabilities.

Customers of Amazon Web Services (AWS) and Informatica can now deploy Informatica Big Data Management in the AWS public cloud leveraging Amazon EMR and other leading Hadoop distributions (e.g. Cloudera, Hortonworks, MapR).

Using Informatica Big Data Management on Amazon EMR provides the following benefits:

- **Faster time to insight.** Dynamic big data integration delivers high throughput data ingestion and data delivery from nearly any source, leveraging Amazon EMR for high performance data processing at scale, delivering the right analytical data to business stakeholders.
- **Faster time to deployment.** The Simple One-Click Automated of Informatica Big Data Management on EMR deployment from the AWS Marketplace allows organizations to quickly and efficiently deploy a big data integration solution on a high performance cloud infrastructure platform.
- **Accelerates data architecture modernization.** If you are planning to modernize your data strategy initiatives on AWS, BDM's rich functionalities such as metadata driven data integration, dynamic mappings and SQL to Mapping conversion, will help you to shorten development cycles and reduce time to market.
- **Delivers clean, complete and trusted data.** Whether you are offloading or extending on-premise applications to the cloud or fully embracing the cloud, collaborative data quality ensures confidence in data fidelity while facilitating data sharing empowering business stakeholders to curate data, audit data holistically, and relate data at scale. Informatica Big Data Management empowers organizations with complete, high-quality, actionable data.

AWS Overview

Amazon Web Services offers the basic building blocks of storage, networking and compute, as well as services such as a managed database, big data, and messaging services. Informatica Big Data Management deployment on EMR can use the following service offerings:

Amazon Elastic Compute Cloud (EC2)

Amazon Elastic Compute Cloud (Amazon EC2) provides scalable computing capacity in the Amazon Web Services (AWS) cloud. Using Amazon EC2 eliminates your need to invest in hardware up front, so you can develop and deploy applications faster. You can use Amazon EC2 to launch as many or as few virtual servers as you need, configure security and networking, and manage storage. Amazon EC2 enables you to scale up or down to handle changes in requirements or spikes in popularity, reducing your need to forecast traffic. Informatica Big Data Management can be deployed on Amazon EC2 infrastructure with the ability to scale up and scale down the environment based on the requirement with zero investment on hardware. Informatica Big Data Management can be deployed in a mixed environment that contains on-premise machines and Amazon EC2 instances.

Amazon Simple Storage Service (S3)

Amazon Simple Storage Service (Amazon S3) provides developers and IT teams with secure, durable, highly-scalable cloud storage. Amazon S3 is easy to use object storage, with a simple web service interface to store and retrieve any amount of data from anywhere on the web. Informatica Big Data Management provides native, high-volume connectivity to Amazon S3 and support for Hive on S3. It is designed and optimized for big data integration between cloud and on-premise data sources to S3 as object stores.

Amazon Redshift

Amazon Redshift is a cloud-based, fast, fully managed, petabyte-scale data warehouse that makes it simple and cost-effective to analyze all data using existing business intelligence tools. Informatica's PowerExchange for Amazon Redshift connector allow users to securely read data from or write data to Amazon Redshift.

Amazon Relational Database Service (RDS)

Amazon Relational Database Service (Amazon RDS) makes it easy to set up, operate, and scale a relational database in the cloud. It provides cost-efficient and resizable capacity while managing time-consuming database administration tasks, freeing you up to focus on your applications and business. Amazon RDS provides you with several familiar database engines to choose from, including Amazon Aurora, Oracle, Microsoft SQL Server, PostgreSQL, MySQL, and MariaDB.

Amazon Aurora

Amazon Aurora is a MySQL-compatible relational database engine that combines the speed and availability of high-end commercial databases with the simplicity and cost-effectiveness of open source databases.

AWS Direct Connect

AWS Direct Connect makes it easy to establish a dedicated network connection from your premises to AWS. Using AWS Direct Connect, you can establish private connectivity between AWS and your datacenter, office, or colocation environment, which in many cases can reduce your network costs, increase bandwidth throughput, and provide a more consistent network experience than Internet-based connections.

Amazon Virtual Private Cloud

Amazon Virtual Private Cloud (Amazon VPC) lets you provision a logically isolated section of the Amazon Web Services (AWS) cloud where you can launch AWS resources in a virtual network that you define. You have complete control over your virtual networking environment, including selection of your own IP address range, creation of subnets, and configuration of route tables and network gateways.

You can easily customize the network configuration for your Amazon Virtual Private Cloud. For example, you can create a public-facing subnet for your webservers that has access to the Internet, and place your backend systems such as databases or application servers in a private facing subnet with no Internet access. You can leverage multiple layers of security, including security groups and network access control lists, to help control access to Amazon EC2 & EMR instances in each subnet.

Regions and Availability Zones

Regions are self-contained geographical locations where AWS services are deployed. Regions have their own deployment of each service. Each service within a region has its own endpoint that you can interact with to use the service.

Regions contain availability zones, which are isolated fault domain within a general geographical location. Some regions have more availability zones than others. While provisioning, you can choose specific availability zones or let AWS select for you.

Networking, Connectivity and Security

Virtual Private Cloud (VPC)

VPC has several different configuration options. See the VPC documentation for a detailed explanation of the options and choose based on your networking requirements. You can deploy Informatica BDM in either public or private subnets.

Connectivity to the Internet and Other AWS Services

Deploying the instances in a public subnet allows them to have access to the Internet for outgoing traffic as well as to other AWS services, such as S3 and RDS.

Private Data Center Connectivity

You can establish connectivity between your datacenter and the VPC hosting your Informatica services by using a VPN or Direct Connect. We recommend using Direct Connect so that there is a dedicated link between the two networks with lower latency, higher bandwidth, and enhanced security. You can also connect to EC2 through the Internet via VPN tunnel if you prefer.

Security Groups

Security Groups are analogous to firewalls. You can define rules for EC2 instances and define allowable traffic, IP addresses, and port ranges. Instances can belong to multiple security groups.

Understanding Amazon EMR Architecture

For more information visit http://docs.aws.amazon.com/ElasticMapReduce/

The central component of Amazon EMR is the cluster. A cluster is a collection of Amazon Elastic Compute Cloud (Amazon EC2) instances. Each instance in the cluster is called a node. Each node has a role within the cluster, referred to as the node type. Amazon EMR also installs different software components on each node type, giving each node a role in a distributed application like Apache Hadoop.

The node types in Amazon EMR are as follows:

Master Node: A node that manages the cluster by running software components which coordinate the distribution of data and tasks among other nodes—collectively referred to as slave nodes—for processing. The master node tracks the status of tasks and monitors the health of the cluster.

Core Node: A slave node that has software components which run tasks and store data in the Hadoop Distributed File System (HDFS) on your cluster.

Task Node: a slave node that has software components which only run tasks. Task nodes are optional.

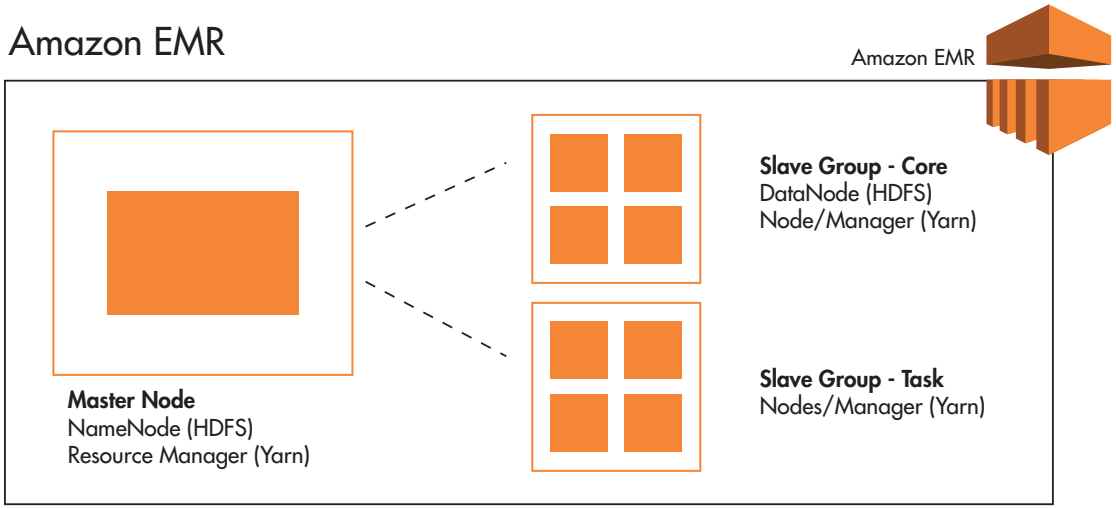

Figure 10. Amazon EMR Architecture.

Amazon EMR service architecture consists of several layers, each of which provides certain capabilities and functionality to the cluster. This section provides an overview of the layers and the components of each.

Storage

The storage layer includes the different file systems that are used with your cluster. There are several different types of storage options as follows.

Hadoop Distributed File System (HDFS)

Hadoop Distributed File System (HDFS) is a distributed, scalable file system for Hadoop. HDFS distributes the data it stores across instances in the cluster, storing multiple copies of data on different instances to ensure that no data is lost if an individual instance fails. HDFS is ephemeral storage that is reclaimed when you terminate a cluster. HDFS is useful for caching intermediate results during MapReduce processing or for workloads which have significant random I/O.

EMR File System (EMRFS)

Using the EMR File System (EMRFS), Amazon EMR extends Hadoop to add the ability to directly access data stored in Amazon S3 as if it were a file system like HDFS. You can use either HDFS or Amazon S3 as the file system in your cluster. Most often, Amazon S3 is used to store input and output data and intermediate results are stored in HDFS.

Local File System

The local file system refers to a locally connected disk. When you create a Hadoop cluster, each node is created from an Amazon EC2 instance that comes with a preconfigured block of preattached disk storage called an instance store. Data on instance store volumes persists only during the lifecycle of its Amazon EC2 instance.

Cluster Resource Management

The resource management layer is responsible for managing cluster resources and scheduling the jobs for processing data.

By default, Amazon EMR uses YARN (Yet Another Resource Negotiator), which is a component introduced in Apache Hadoop 2.0 to centrally manage cluster resources for multiple data-processing frameworks. However, there are other frameworks and applications that are offered in Amazon EMR that do not use YARN as a resource manager. Amazon EMR also has an agent on each node which administers YARN components, keeps the cluster healthy, and communicates with the Amazon EMR service.

Data Processing Frameworks

The data processing framework layer is the engine used to process and analyze data. There are many frameworks available that run on YARN or have their own resource management. Different frameworks are available for different kinds of processing needs, such as batch, interactive, in-memory, streaming, and so on. The framework that you choose depends on your use case. This impacts the languages and interfaces available from the application layer, which is the layer used to interact with the data you want to process. The main processing frameworks available for Amazon EMR are Hadoop MapReduce and Spark.

Hadoop MapReduce

Hadoop MapReduce is an open-source programming model for distributed computing. It simplifies the process of writing parallel distributed applications by handling all of the logic, while you provide the Map and Reduce functions. The Map function maps data to sets of key value pairs called intermediate results. The Reduce function combines the intermediate results, applies additional algorithms, and produces the final output. There are multiple frameworks available for MapReduce, such as Hive, which automatically generate Map and Reduce programs.

Apache Spark

Spark is a cluster framework and programming model for processing big data workloads. Like Hadoop MapReduce, Spark is an open-source, distributed processing system but uses directed acyclic graphs for execution plans and leverages in-memory caching for datasets. When you run Spark on Amazon EMR, you can use EMRFS to directly access your data in Amazon S3. Spark supports multiple interactive query modules such as SparkSQL

Informatica Blaze

Informatica Blaze is the industry's unique data processing engine integrated with YARN to provide intelligent data pipelining, job partitioning, job recovery, and scalability, which is optimized to deliver high performance, scalable data processing leveraging Informatica's cluster aware data integration technology.

Informatica Big Data Management Deployment Options

Informatica Big Data Management can be deployed on AWS EMR or on-premise. The table below describes the benefits of deploying on-premise vs. Amazon EMR:

	BDM ON-PREMISE	BDM ON EMR
Scalability	• Horizontal scalability involves lead time to procure physical servers. • Vertical scalability often involves guesswork to predict future load.	• Scale up in minutes, not weeks. • The infrastructure is easy to configure and can be highly automated. If your needs change, simply scale back and only pay for what you use.
Point in time snapshots	• Managing snapshots in an on- premise environment can be costly and complex.	• Easily automate your backup strategy and only pay for what you use, when you use it.
Back-up strategy	• An often costly effort that involves multiple vendors and media, with different management planes.	• Back up your data to S3 for a durable, low cost approach and utilize the built-in data lifecycle policies to get the right storage at the right price.
Amazon RDS and Redshift connectivity	• Connecting to AWS services through the corporate firewall adds complexity and latency.	• Leverage a secure, low latency connection to popular services like Redshift.

Deployment Architecture for Big Data Management on Amazon EMR

Deployment Topology

Informatica Big Data Management can be deployed on Amazon EMR in three ways:

1. Simple One-Click Automated BDM-EMR Deployment from AWS Marketplace
2. Deploying BDM on an Existing Amazon EMR Instance with BDM Service on-premise
3. Deploying BDM on an Existing Amazon EMR Instance with BDM Service in the cloud

This section describes the first option, Simple One-Click Automated BDM-EMR Deployment from AWS Marketplace.

To deploy BDM using option 2 or 3 please refer to the Informatica Big Data Management Installation & Configuration for Amazon EMR on the Informatica Network: https://kb.informatica.com/howto/6/Pages/18/495239.aspx

Simple One-Click Automated BDM-EMR Deployment from AWS Marketplace

Informatica and Amazon AWS provide the availability of a completely automated deployment of Informatica Big Data Management on Amazon EMR cluster thru the Amazon Marketplace. By default, the deployment consists of a minimum recommended M3 instance type using the m3.xlarge model that provides a balance of compute, memory, and network resources required for Informatica BDM services and Amazon EMR cluster. The user has the option to choose other M3 instance types or C3 Compute-optimized instances for high performing processing and can choose any number of core nodes required for the Amazon EMR Cluster.

The automated process assigns the correct number of BDM Server and Amazon EMR nodes and provisions the nodes with the latest Informatica Big Data Management software and the Amazon EMR cluster with the following services: HDFS, EMRFS, YARN, Hive, and Spark. Figure 10 shows the deployment flow using Amazon AWS Marketplace to launch Informatica BDM on Amazon EMR.

One-click deployment is best suited for proof of concept and prototyping big data projects where Amazon AWS and Informatica Big Data Management services are deployed automatically on AWS infrastructure.

To provision an Amazon EMR cluster with Informatica Big Data Management using the simple one-click automation script, follow the instructions here: How To: BDM ON EMR.

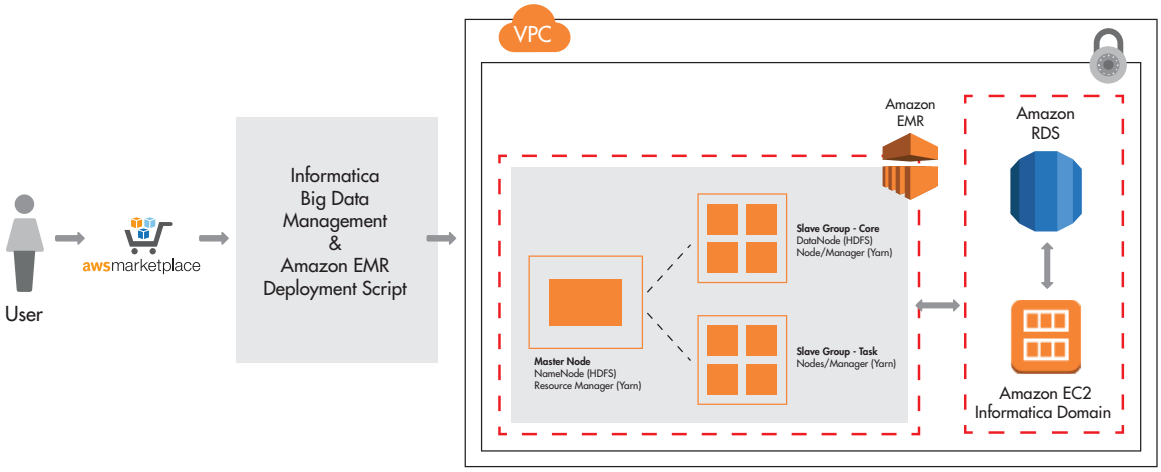

Figure 11. Amazon EMR Architecture.

Informatica Services on Amazon AWS

Informatica BDM is installed and configured on an Amazon EC2 instance during the provision of the nodes from the one-click deployment. The deployment automatically creates the following Informatica services: Domain, Model Repository, Data Integration service and assigns the connection to the Amazon EMR cluster for HDFS, Hive.

The Informatica domain and repository database are hosted on Amazon RDS using MS SQL Server which handles management tasks, such as backups, patch management, and replication. To access the Informatica Services, the Informatica client must be installed on a Microsoft Windows machine.

The following figure shows Informatica Services running on Amazon AWS:

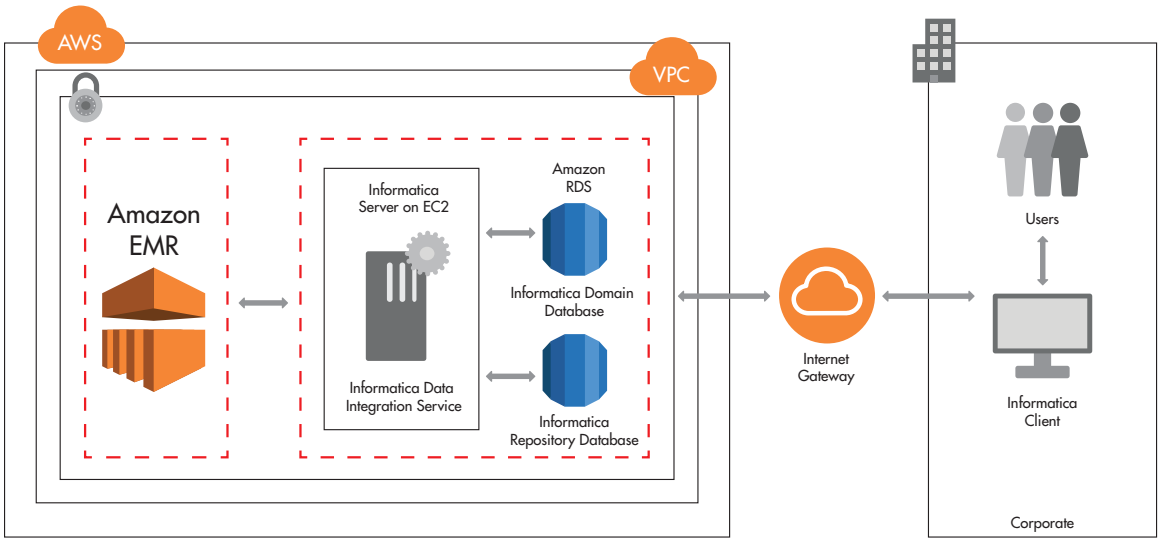

Figure 12. Informatica BDM on EC2.

Understanding Informatica Architecture

The following describes the information services that are configured during the one-click deployment:

- Informatica Domain is the fundamental administrative unit of the Informatica Platform. The Informatica Platform has a service-oriented architecture that provides the ability to scale services and share resources across multiple machines. High availability functionality helps minimize service downtime due to unexpected failures or scheduled maintenance in the Informatica environment.
- Informatica Node is a logical representation of a physical or virtual machine in the Informatica Domain. Each Informatica Node in the Informatica Domain runs application services, such as the Informatica Model Repository and Informatica Data Integration Service.
- Informatica Model Repository Service manages the Model Repository which is a relational database that stores all the metadata for projects created using Informatica Client tools. The Model repository also stores run-time and configuration information for applications that are deployed to a Data Integration Service.
- Informatica Data Integration Service is a compute component within the Informatica domain that manages requests to submit big data integration, big data quality and profiling jobs to the Hadoop cluster for processing.
- Informatica Client, specifically the Informatica developer tool allows data engineers to design and implement big data integration, big data quality and profiling solution that execute on the Hadoop cluster.

The Informatica Domain & Informatica Model Repository databases are configured on Amazon RDS using Microsoft SQL Server.

Understanding Informatica Support for Big Data Processing Frameworks

As discussed earlier, big data processing frameworks have the ability to process any data at any latency, batch, near real-time and real-time. Informatica's Big Data Management solutions support multiple processing paradigms, such as MapReduce, Hive on Tez, Informatica Blaze, Spark to execute each workload on the best possible processing engine.

Informatica Blaze, a native YARN application, is a distributed processing engine with the ability to scale and perform high speed data processing of large complex batch workloads via a natively embedded Informatica data transformation engine on Hadoop.

Using an in-built Smart Executor, the Informatica Data Integration Service, automatically determines the optimal batch processing layer at runtime across a variety of execution engines such as Informatica Blaze, Hive on Tez, Hive on MapReduce and Apache Spark on Hadoop or native Informatica server.

Storage Segregation

Informatica services stores a variety of types of data when installed on EC2: workflow and session logs, repository backups, and both persistent and non-persistent cache files generated by transformations. For maximum performance on EC2, use EBS volumes for persistent data and ephemeral instance storage for temporary cache data. Examples of data suitable for EBS include: $INFA_HOME, repository backup, and the session log directory.

Network Prerequisites

The Informatica client installed on an on-premise Microsoft Windows machine, at least 32Mbps of sustained bandwidth is recommended. If your instance supports an EBS-optimized flag, enable it to add up to 500 mbps of bandwidth to EBS. The amount of bandwidth available depends on type of instance chosen.

Security Group Setting

The following are provisioned within the security group to allow traffic on the following ports:

Port Name	Default Port Numbers
Node Port	6005
Service Manager Port	6006
Service Manager Shutdown Port	6007
Informatica Administrator Port	HTTP: 6008 HTTPS: 8443
Informatica Administrator Shutdown Port	6009

For more information about Informatica port administration, see

https://kb.informatica.com/h2l/HowTo%20Library/1/0519-Informatica Port Adminstration-H2L.pdf

Understanding Amazon EMR Cluster Types & Architectures for Informatica Big Data Management

Amazon EMR can be deployed using a variety of configurations of architectures, each with its own advantages and disadvantages. This section describes the Amazon EMR deployment types and architectures in conjunction with Informatica Big Data Management.

For more information about Amazon EMR best practices, see https://d0.awsstatic.com/whitepapers/aws-amazon-emr-best-practices.pdf

Amazon EMR provides two methods to configure a cluster: **transient** and **persistent**. Transient clusters are shut down when the jobs are complete. For example, if a batch-processing job that pulls weblogs from Amazon S3 and processes the data once a day, it is more cost effective to use transient clusters to process weblog data and shut down the nodes when the processing is complete. Persistent clusters continue to run after data processing is complete. An Infrastructure Architect needs to consider which configuration method works best for the organizations use case as both have advantages and disadvantages. Please refer to the Amazon EMR best practices noted above. Informatica Big Data Management supports both cluster types.

Common Amazon EMR Architectures Patterns for Informatica Big Data Management

A common challenge of processing large amounts of data in Hadoop on the cloud is moving data from origin to processing infrastructure to the final destination. The tenant for Hadoop is the ability to process and store large data sets in a scalable, reliable distributed computing environment.

Amazon AWS provides a number of methods for moving large amounts of data from source to Amazon S3 or from Amazon S3 to Amazon EMR and the Hadoop Distributed File System (HDFS). Proper management of the data bandwidth is crucial to the success of big data processing. Optimized networks can reduce network costs, increase bandwidth throughput, and provide a more consistent network experience.

Described below are patterns Informatica Big Data Management supports to move data to an Amazon EMR cluster for processing.

Pattern 1: Using Amazon S3

In this first pattern, data is loaded to Amazon S3 using Informatica Big Data Management (BDM) using PowerExchange for Amazon S3 connectivity. For data processing, Informatica BDM mapping logic pulls data from Amazon S3 and sends it for processing to Amazon EMR. Amazon EMR does not copy the data to local disk or HDFS; instead the mappers open multithreaded HTTP connections to Amazon S3, pulls data to the Amazon EMR cluster and process data in streams.

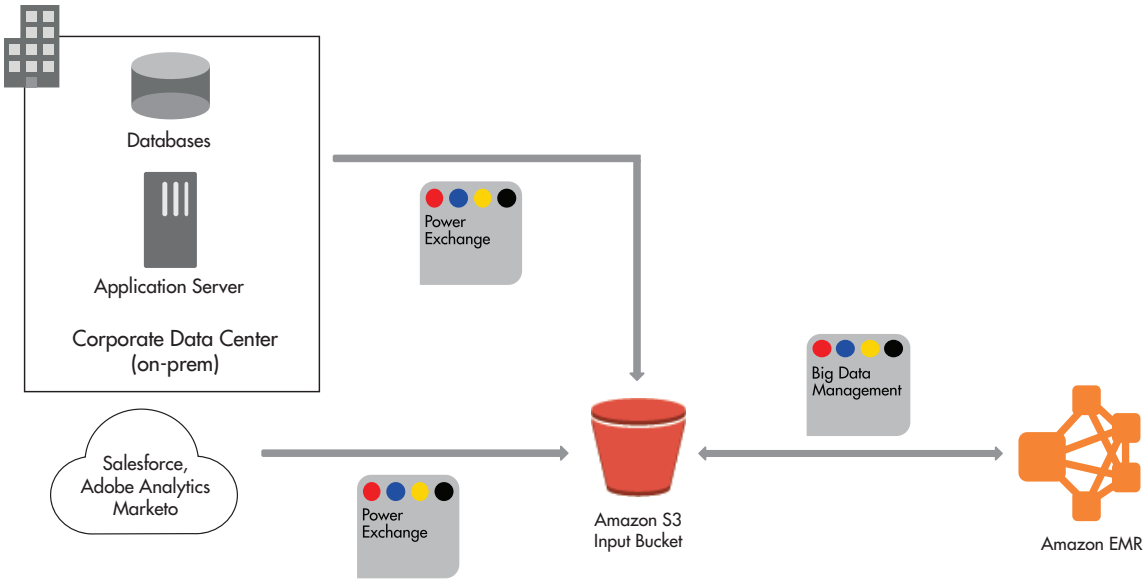

Figure 13. Pattern 1 using Amazon S3.

Pattern 2: Using HDFS and Amazon S3 as Backup Storage

In this pattern, Informatica Big Data Management writes data directly to HDFS and leveraging a persistent Amazon EMR cluster to process the data and periodically copy data to Amazon S3 as the backup storage. The advantage of this pattern is the ability to process data without copying it to Amazon EMR. Even though this method may improve performance the disadvantage is durability. Since Amazon EMR uses ephemeral disk to store data, data could be lost if the Amazon EMR EC2 instance fails. HDFS replicates data within the Amazon EMR cluster and can usually recover from node failures. However, data loss could still occur if the number of lost nodes is greater than your replication factor. It is highly recommended to back up HDFS data to Amazon S3 periodically.

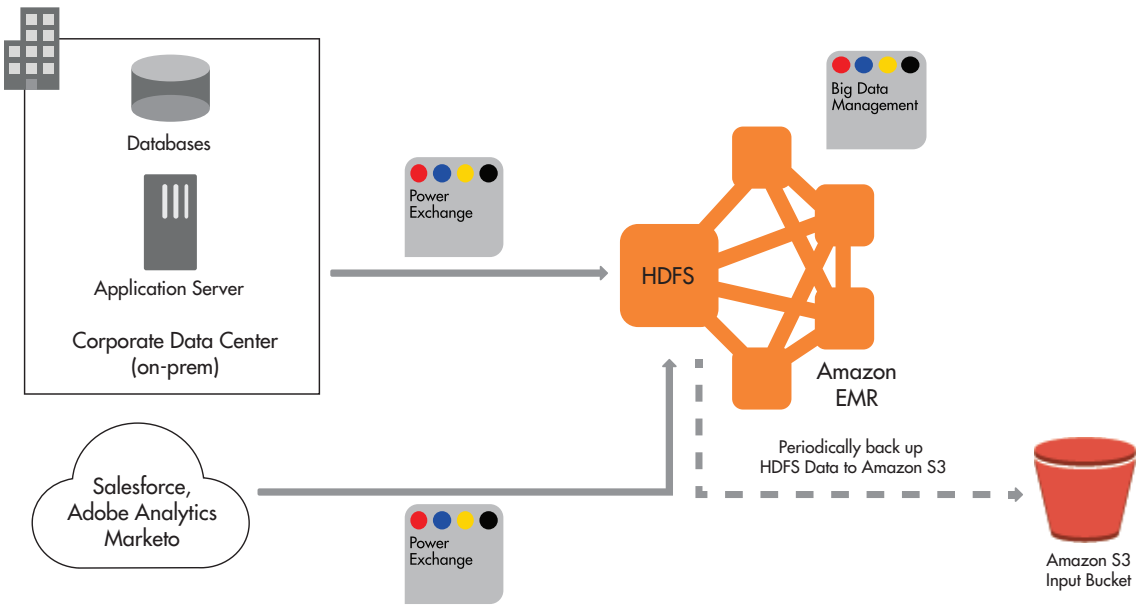

Figure 14. Pattern 2 using HDFS and Amazon S3 as Backup.

Leveraging Amazon EMR for Informatica's Marketing Data Lake Solution

The following architecture diagram describes the process flow for using Informatica Big Data Management on Amazon EMR as it relates to this reference architecture. The diagram below illustrates the data flow process using Informatica BDM and Amazon EMR, Amazon S3 and Amazon Redshift. The process follows pattern 1 discussed in the earlier section. Identified are five fundamental processes that an Enterprise Architect must consider when architecting a Marketing Data Lake solution.

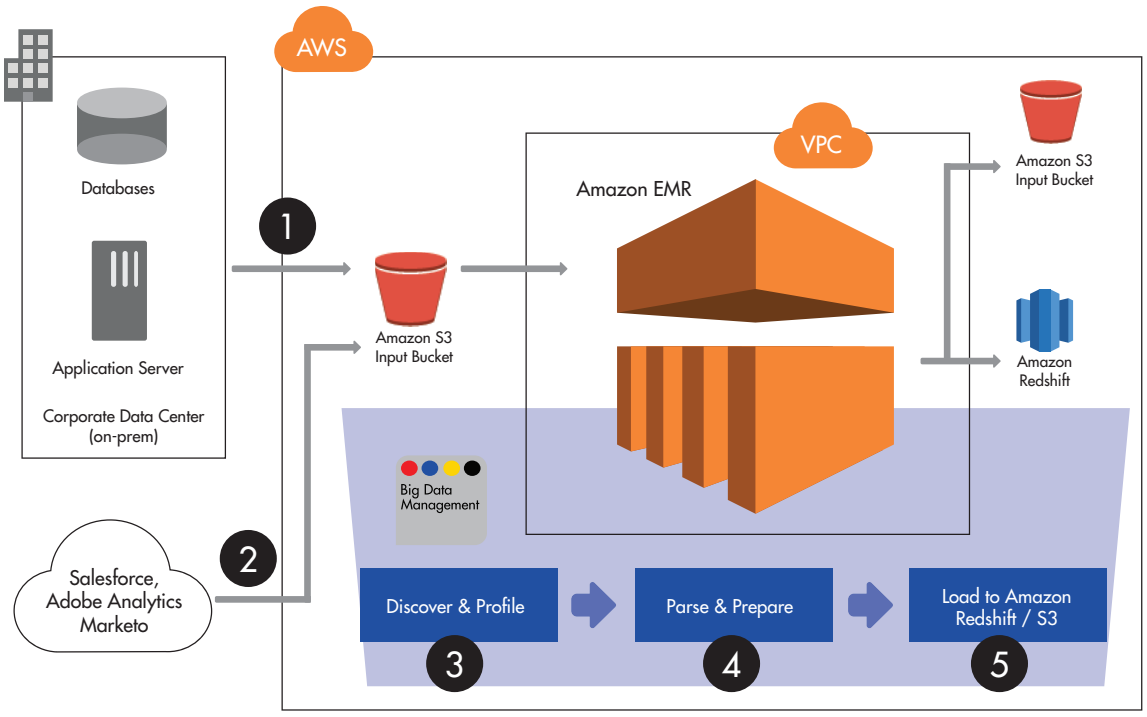

Figure 15. Informatica BDM Process Flow using Amazon EMR.

1. Offload infrequently used data & batch load raw data onto a defined landing zone in an Amazon S3 bucket. As discussed in the section, "Data Lake Management Reference Architecture for the Marketing Data Lake", marketing data stored on data warehouses or an application server can be offloaded to a dedicated area freeing up space in the current Enterprise Data Warehouse. Instead of feeding data from source systems into the warehouse, raw transactional and multi-structured data is loaded directly onto Amazon S3, further reducing impact on the warehouse.

2. Collect and stream real-time machine and sensor data. Data generated by machines and sensors, including application and weblog files, can be collected in real time and streamed directly into Amazon S3 instead of being staged in a temporary file system or the data warehouse.

3. Discover and profile data stored on Amazon S3. Data can be profiled to better understand its structure and context and adding requirements for enterprise accountability, control, and governance for compliance with corporate and governmental regulations and business SLAs.

About Informatica

Informatica is 100 percent focused on data because the world runs on data. Organizations need business solutions around data for the cloud, big data, real-time and streaming. Informatica is the world's No. 1 provider of data management solutions, in the cloud, on-premise or in a hybrid environment. More than 7,000 organizations around the world turn to Informatica for data solutions that power their businesses.

4. Parse and prepare data from weblogs, application server logs or sensor data. Typically these data types are either in multi-structured or unstructured formats which can be parsed to extract features and entities, and data quality techniques can be applied. Prebuilt transformations and data quality and matching rules can be executed natively in Amazon EMR, preparing data for analysis.
5. After data has been cleansed and transformed using Amazon EMR, high-value curated data can be moved from EMR to Amazon S3 output bucket or Amazon Redshift where data is directly accessible by the BI reports, applications, and users.

Summary

The Marketing Data Lake can consolidate data across all marketing platforms which allows for clear insight into product interest, understanding customer behavior, and improving marketing operations.

The Informatica Data Lake Management reference architecture for the marketing data lake provides enterprises a roadmap for foundational capabilities necessary to derive value from big data.

Informatica and Amazon AWS allow customers to easily deploy Informatica Big Data Management on Amazon AWS leveraging Amazon AWS's flexible, scalable, reliable and secure infrastructure to build next generation marketing platform, the Marketing Data Lake.

For more information, visit www.informatica.com/bigdata.

Worldwide Headquarters 2100 Seaport Blvd., Redwood City, CA 94063, USA Phone: 650.385.5000, Toll-free in the US: 1.800.653.3871
www.informatica.com linkedin.com/company/informatica twitter.com/Informatica

IN09_1116_03273_0218