



Siperian Hub

Overview

Copyright 2009 Siperian, Inc.

Copyright 2009 Siperian Inc. [Unpublished - rights reserved under the Copyright Laws of the United States]

Siperian and the Siperian logo are trademarks or registered trademarks of Siperian, Inc. in the US and other countries. All other products or services mentioned are the trademarks or service marks of their respective companies or organizations.

THIS DOCUMENTATION CONTAINS CONFIDENTIAL INFORMATION AND TRADE SECRETS OF SIPERIAN, INC. USE, DISCLOSURE OR REPRODUCTION IS PROHIBITED WITHOUT THE PRIOR EXPRESS WRITTEN PERMISSION OF SIPERIAN, INC.

Updated: November, 2009

Contents

Preface

Intended Audience	vii
Organization	viii
Learning About Siperian Hub	viii
Contacting Siperian	xi

Chapter 1: Introduction to Siperian Hub

Master Data Management	2
Master Data and Master Data Management	2
Customer Case Studies	3
Key Adoption Drivers for Master Data Management.....	4
Siperian Hub as the Enterprise MDM Platform.....	5
About Siperian Hub	5
Core Capabilities	6

Chapter 2: Siperian Hub Architecture

Key Siperian Hub Components	10
Core Components	11
Hub Store.....	11
Hub Server.....	12
Cleanse Match Server.....	13
Hub Console	14
Master Reference Manager.....	15
Hierarchy Manager	16
Activity Manager.....	17
Security Access Manager	18
Metadata Manager	19

Services Integration Framework.....	20
Business Data Director	21

Chapter 3: Key Concepts

Inbound and Outbound Data Flows	24
Main Inbound Data Flow (Reconciliation).....	24
Main Outbound Data Flow (Distribution)	26
Batch and Real-Time Processing.....	27
Batch Process Flow.....	27
Real-time Process Flow.....	28
Batch Processing	29
Land Process.....	31
Stage Process.....	32
Load Process.....	33
Tokenize Process	34
Match Process.....	35
Consolidate Process.....	36
Publish Process.....	37
Real-Time Processing.....	37
Databases in the Hub Store.....	38
Content Metadata.....	39
Base Objects	39
Cross-Reference (XREF) Tables	39
History Tables.....	40
Workflow Integration and State Management	40
Hierarchy Management	41
Relationships.....	41
Hierarchies	41
Entities.....	41
Activity Management.....	42
Rules and Actions	42
Remote Packages.....	42

Chapter 4: Topics for Siperian Hub Users

Administrators.....	44
About Siperian Hub Administrators.....	44
Documentation Resources for Siperian Hub Administrators	44
Developers.....	45
About Siperian Hub Developers.....	45
Documentation Resources for Siperian Hub Developers.....	45
Data Stewards.....	46
About Siperian Hub Data Stewards.....	46
Documentation Resources for Siperian Hub Data Stewards	46

Preface

Welcome to the Siperian Hub™ *Overview*. This document provides an overview of the Siperian Hub suite of products, describes the product architecture, and defines key concepts that you need to understand in order to use Siperian Hub in your organization.

Intended Audience

This document is intended to introduce important Siperian Hub concepts to anyone who is involved in a Siperian Hub implementation. This document is primarily directed at those who are charged with the responsibility of managing, implementing, or using Siperian Hub in an organization. Its audience includes—but is not limited to—project managers, installers, developers, administrators, system integrators, database administrators, data stewards, and other technical specialists associated with a Siperian Hub implementation. The goal of this document is to provide users with a succinct but comprehensive, high-level understanding of the product suite, along with instructions on where to go in the product documentation set to find more information about specific topics.

Organization

This guide contains the following chapters:

Chapter 1, “Introduction to Siperian Hub”	Introduces Siperian Hub as the premier enterprise platform for Master Data Management (MDM).
Chapter 2, “Siperian Hub Architecture”	Describes the overall Siperian Hub architecture and describes its key constituent components.
Chapter 3, “Key Concepts”	Describes key concepts that every Siperian Hub user should understand.
Chapter 4, “Topics for Siperian Hub Users”	Describes topics of interest for various types of Siperian Hub users.

Learning About Siperian Hub

What’s New in Siperian Hub

What’s New in Siperian Hub describes the new features in this Siperian Hub release.

Siperian Hub Release Notes

The *Siperian Hub Release Notes* contain important information about this Siperian Hub release. Installers should read the *Siperian Hub Release Notes* before installing Siperian Hub.

Siperian Hub Overview

The *Siperian Hub Overview* introduces Siperian Hub, describes the product architecture, and explains core concepts that users need to understand before using the product. All users should read the *Siperian Hub Overview* first.

Siperian Hub Installation Guide

The *Siperian Hub Installation Guide* explains to installers how to set up Siperian Hub, the Hub Store, Cleanse Match Servers, and other components. There is a *Siperian Hub Installation Guide* for each supported platform.

Siperian Hub Upgrade Guide

The *Siperian Hub Upgrade Guide* explains to installers how to upgrade a previous Siperian Hub version to the most recent version.

Siperian Hub Cleanse Adapter Guide

The *Siperian Hub Cleanse Adapter Guide* explains to installers how to configure Siperian Hub to use the supported adapters and cleanse engines.

Siperian Hub Data Steward Guide

The *Siperian Hub Data Steward Guide* explains to data stewards how to use Siperian Hub tools to consolidate and manage their organization's data. Data stewards should read the *Siperian Hub Data Steward Guide* after having reading the *Siperian Hub Overview*.

Siperian Hub Administrator Guide

The *Siperian Hub Administrator Guide* explains to administrators how to use Siperian Hub tools to build their organization's data model, configure and execute Siperian Hub data management processes, set up security, provide for external application access to Siperian Hub services, and other customization tasks. Administrators should read the *Siperian Hub Administrator Guide* after having reading the *Siperian Hub Overview*.

Siperian Hub Services Integration Framework Guide

The *Siperian Hub Services Integration Framework Guide* explains to developers how to use the Siperian Hub Services Integration Framework (SIF) to integrate Siperian Hub functionality with their applications, and how to create applications using the data provided by Siperian Hub. SIF allows developers to integrate Siperian Hub smoothly with their organization's applications. Developers should read the *Siperian Hub Services Integration Framework Guide* after having reading the *Siperian Hub Overview*.

Siperian Hub Metadata Manager Guide

The *Siperian Hub Metadata Manager Guide* explains how to use the Siperian Hub Metadata Manager tool to validate their organization's metadata, promote changes between repositories, import objects into repositories, export repositories, and related tasks.

Siperian Hub Resource Kit User Guide

The *Siperian Hub Resource Kit User Guide* explains how to install and use the Siperian Hub Resource Kit, which is a set of utilities, examples, and libraries that assist developers with integrating the Siperian Hub into their applications and workflows. This document also provides a description of the various sample applications that are included with the Resource Kit.

Siperian Training and Materials

Siperian provides live, instructor-based training to help professionals become proficient users as quickly as possible. From initial installation onward, a dedicated team of qualified trainers ensure that an organization's staff is equipped to take advantage of this powerful platform. To inquire about training classes or to find out where and when the next training session is offered, please visit Siperian's web site (www.siperian.com) or contact Siperian directly.

Contacting Siperian

Technical support is available to answer your questions and to help you with any problems encountered using Siperian products. Please contact your local Siperian representative or distributor as specified in your support agreement. If you have a current Siperian Support Agreement, you can contact Siperian Technical Support:

Method	Contact Information
World Wide Web	http://www.siperian.com
Email	support@siperian.com
Voice	U.S.: 1-866-SIPERIAN (747-3742)

We are interested in hearing your comments about this book. Send your comments to:

by Email: docs@siperian.com

by Postal Service: Documentation Manager
Siperian, Inc.
100 Foster City Blvd.
2nd Floor
Foster City, California 94404 USA

Introduction to Siperian Hub

This chapter introduces master data management (MDM) as it relates to the Siperian Hub suite of products.

Chapter Contents

- [Master Data Management](#)
- [Siperian Hub as the Enterprise MDM Platform](#)

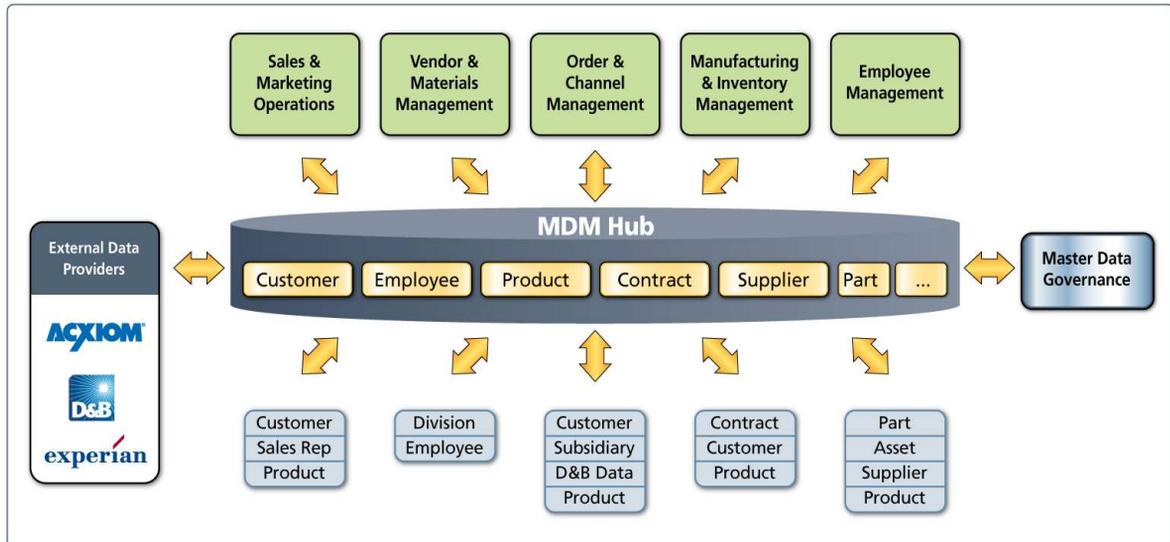
Master Data Management

This section introduces master data management as a discipline for improving data reliability across the enterprise.

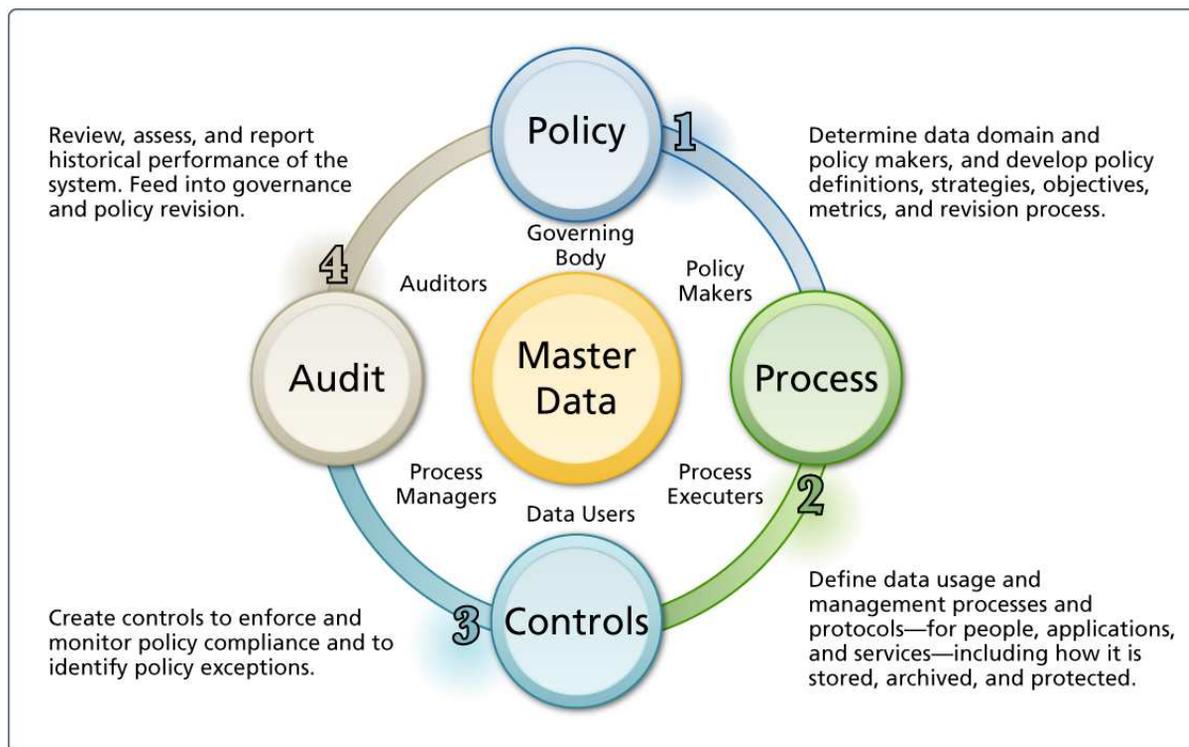
Master Data and Master Data Management

Master data is a collection of common, core entities—along with their attributes and their values—that are considered critical to a company's business, and that are required for use in two or more systems or business processes. Examples of master data include customer, product, employee, supplier, and location data. Complexity arises from the fact that master data is often strewn across many channels and applications within an organization, invariably containing duplicate and conflicting data.

Master Data Management (MDM) is the controlled process by which the master data is created and maintained as the *system of record* for the enterprise. MDM is implemented in order to ensure that the master data is validated as correct, consistent, and complete. Optionally, MDM can be implemented to ensure that Master Data is circulated in context for consumption by internal or external business processes, applications, or users.



Ultimately, MDM is deployed as part of the broader Data Governance program that involves a combination of technology, people, policy, and process.

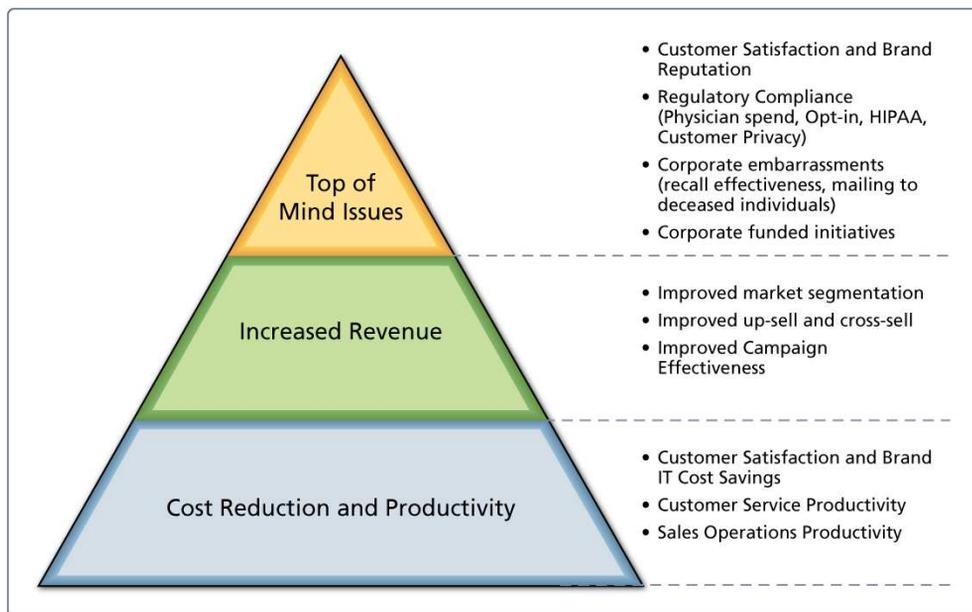


Organizations are implementing master data management solutions to enhance data reliability and data maintenance procedures. Tight controls over data imply a clear understanding of the myriad data entities that exist across the organization, data maintenance processes and best practices, and secure access to the usage of data.

Customer Case Studies

The Siperian web site (<http://www.siperian.com/>) provides case studies that describe how Siperian customers have benefited by deploying Siperian Hub in their organizations.

Key Adoption Drivers for Master Data Management



Organizations are implementing master data management solutions to achieve:

- **Regulatory compliance**, such as financial reporting and data privacy requirements.
- **Cost savings** by streamlining business processes, consolidating software licenses, and reducing the costs associated with data administration, application development, data cleansing, third-party data providers, and capital costs.
- **Productivity improvements** across the organization by reducing duplicate, inaccurate, and poor-quality data, helping to refocus resources on more strategic or higher-value activities.
- **Increased revenue** by improving visibility and access to accurate customer data, resulting in increased yields for marketing campaigns and better opportunities for cross-selling and up-selling to customers and prospects.
- **Strategic goals**, such as customer loyalty and retention, supply chain excellence, strategic sourcing and contracting, geographic expansion, and marketing effectiveness.

Siperian Hub as the Enterprise MDM Platform

This section describes Siperian MDM Hub (hereafter referred to as *Siperian Hub*) as an MDM platform.

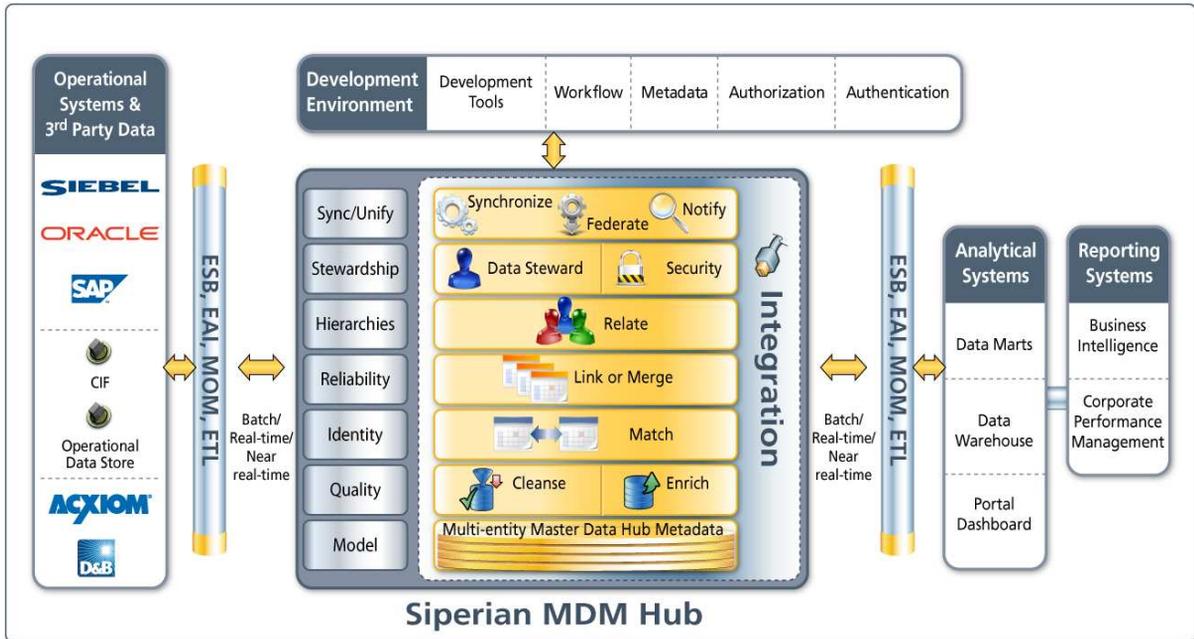
About Siperian Hub

Siperian Hub is the best platform available today for deploying MDM solutions across the enterprise. Siperian Hub offers an integrated, model-driven, and flexible enterprise MDM platform that can be used to create and manage all kinds of master data.

Characteristic	Description
Integrated	Siperian Hub provides a single code-base with all data management technologies, and handles all entity data types in all modes (for operational and analytical use).
Model-Driven	Siperian Hub models an organization's business definitions according to its own requirements and style. All metadata—and business services—are generated on the organization's definitions. Siperian Hub is configurable with history and lineage.
Flexible	Siperian Hub implements all types of MDM styles—registry, reconciled trusted source of truth—and styles can be combine within a single hub. Siperian Hub also coexists with legacy hubs.

Core Capabilities

The following figure shows a functional overview of Siperian Hub’s core capabilities.



As data arrives at the hub, it is often not standardized. This standardization includes name corrections (for example, Mike to Michael), address standardizations (for example, 123 Elm St., NY NY to 123 Elm Street, New York, NY), as well as data transformations (one data model to another). The data can be enriched or augmented with data from third-party data providers such as D&B and Acxiom. Siperian Hub provides out-of-the-box integration with major third-party data providers within its user interface.

After data standardization and enrichment, common records are identified by rapidly matching against each other. Once common records are identified, you can either link them as a registry style or merge the best attributes from the matched records to create the “Best Version of the Truth.” This reconciliation process—achieved within the Siperian Trust Framework and governed by configured business rules—provides the best attributes from contributing systems.

Relating people and organizations is a key requirement for many organizations. Siperian Hub's Hierarchy Management capabilities let users group people into households and companies into corporate hierarchies.

Siperian Hub also provides GUI-based functionality, enabling users to define and configure business rules that affect how data is cleansed, matched, and merged. This data management workflow presents the exceptions or non-automated matches to the data steward for resolution.

All data in the Siperian Hub is available based on the entitlement rules that are put in place, ensuring that only authorized users can view or modify the data and, if necessary, mask important data (such as tax ID numbers).

One common goal of sharing the data in Siperian Hub is to synchronize it with contributing source systems as well as downstream systems. Siperian Hub can be configured to handle these synchronizations in real time, near-real time, or batch mode. If in real time or near-real time mode, Siperian Hub is smart enough to avoid loop backs with the system that initiated the change in the first place.

Siperian Hub also has the ability to dynamically aggregate transaction and activity data into a central record, leveraging federated query technology built into the hub. This allows organizations to store only the reference data in the hub while providing access to all the transaction data in real time.

With the complete view of the client and their transactions, users can configure notification events that are triggered when data changes and can kick off a workflow process, an email, or invoke a web service. This allows organizations to respond to changes as they happen.

Finally, Siperian Hub can be configured to share data using pre-configured web services, or organizations can assemble higher-level functions by orchestrating multiple services.

Siperian Hub Architecture

This chapter describes the Siperian Hub architecture and its constituent components.

Chapter Contents

- [Key Siperian Hub Components](#)
- [Core Components](#)
- [Master Reference Manager](#)
- [Hierarchy Manager](#)
- [Activity Manager](#)
- [Security Access Manager](#)
- [Metadata Manager](#)
- [Services Integration Framework](#)
- [Business Data Director](#)

Key Siperian Hub Components

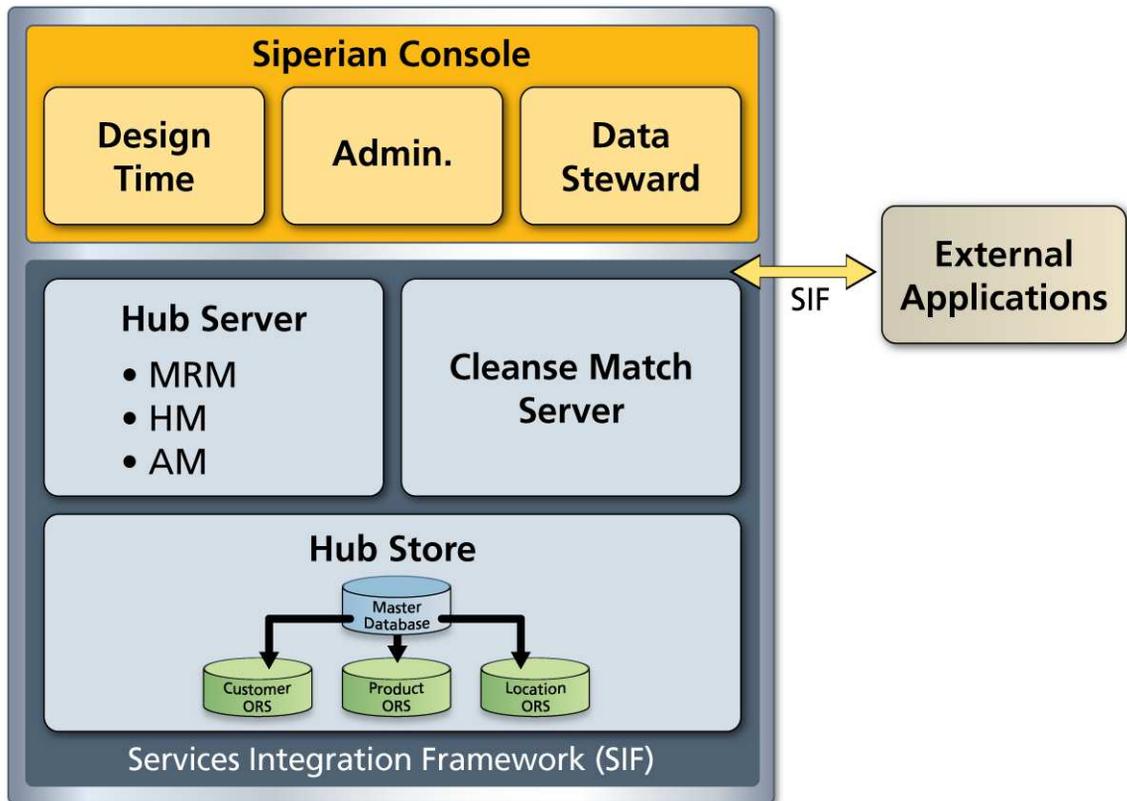
Siperian Hub includes the following key components:

Component	Description
Core Components	Provides core Siperian Hub functionality.
Master Reference Manager™	Manages the data cleansing and provides the matching and consolidating functionality to create the most accurate master records.
Hierarchy Manager™	Builds and manages the data describing the relationships between master records. Also known as <i>HM</i> .
Activity Manager™	Evaluates data events, synchronizes master data, and delivers unified views of reference and activity data from disparate sources. Also known as <i>AM</i> .
Security Access Manager	Provides comprehensive and highly-granular security mechanisms to ensure that only authenticated and authorized users have access to Siperian Hub data, resources, and functionality. Also known as <i>SAM</i> .
Metadata Manager	Allows administrators to manage metadata in their Siperian Hub implementation. Also known as <i>MET</i> .
Services Integration Framework	Enables external applications to request Siperian Hub operations and gain access to Siperian Hub resources via an application programming interface (API). Also known as <i>SIF</i> .
Business Data Director™	Data governance application that enables business users to create, manage, consume, and monitor master data in Siperian Hub. Also known as <i>BDD</i> .

For more information about the components mentioned in this chapter, refer to “[Learning About Siperian Hub](#)” on page viii.

Core Components

The following figure shows the Siperian Hub core components:



Hub Store

The Hub Store is where business data is stored and consolidated. The Hub Store contains common information about all of the databases that are part of a Siperian Hub implementation (as described in [“Databases in the Hub Store”](#) on page 38). The Hub Store resides in a supported database server environment.

The Hub Store contains:

- all the master records for all entities across different source systems
- rich metadata and the associated rules needed to determine and continually maintain only the most reliable cell-level attributes in each master record
- logic for data consolidation functions, such as merging and unmerging data

For more information about the Hub Store, refer to the following documentation.

Task	Topic(s)
Installation	“Installing Hub Store” in the <i>Siperian Hub Installation Guide</i> for your platform
Concepts	“About the Hub Store” in the <i>Siperian Hub Administrator Guide</i>
Configuration	“Configuring Operational Record Stores and Datasources” in the <i>Siperian Hub Administrator Guide</i>

Hub Server

The Hub Server is the run-time component that manages core and common services for the Siperian Hub. The Hub Server is a J2EE application, deployed on the application server, that orchestrates the data processing within the Hub Store, as well as integration with external applications.

For more information about the Hub Server, refer to the following documentation.

Task	Topic(s)
Installation	“Installing Hub Server” in the <i>Siperian Hub Installation Guide</i> for your platform
Concepts	“About the Hub Server” in “Installing Hub Server” in the <i>Siperian Hub Installation Guide</i>
Configuration	“Configuring Hub Server” in “Installing Hub Server” in the <i>Siperian Hub Installation Guide</i> <i>Siperian Hub Administrator Guide</i>

Cleanse Match Server

The Cleanse Match Server run-time component handles cleanse and match requests and is deployed in the application server environment. The Cleanse Match Server contains:

- a cleanse server that handles data cleansing operations
- a match server that handles match operations

The Cleanse Match Server interfaces with any of the supported cleanse engines, as described in *Siperian Hub Cleanse Adapter Guide*. The Cleanse Match Server and the cleanse engine work together to standardize the data and to optimize the data for match and consolidation.

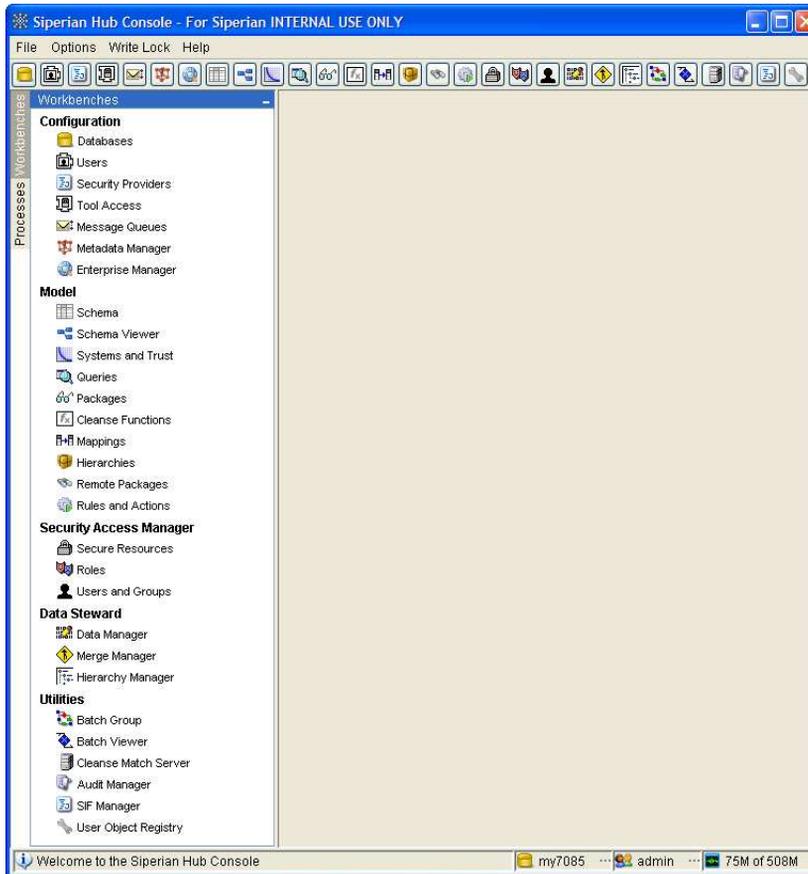
For more information about Cleanse Match Servers, refer to the following documentation.

Task	Topic(s)
Installation	“Installing the Cleanse Match Server” in the <i>Siperian Hub Installation Guide</i> for your platform
Concepts	“About the Cleanse Match Server” in “Installing the Cleanse Match Server” in the <i>Siperian Hub Installation Guide</i> for your platform
Configuration	“Configuring Cleanse Match Servers” in “Configuring Data Cleansing” in the <i>Siperian Hub Administrator Guide</i>

Hub Console

The Hub Console is the Siperian Hub user interface that comprises a set of tools for administrators and data stewards. Each tool allows users to perform a specific action, or a set of related actions, such as building the data model, running batch jobs, configuring the data flow, running batch jobs, configuring external application access to Siperian Hub resources, and other system configuration and operation tasks.

The Hub Console is packaged inside the Hub Server application. It can be launched on any client machine via a URL using a browser and Sun's Java Web Start.



Note: The available tools in the Hub Console depend on your Siperian license agreement. Therefore, your Hub Console tool might differ from the previous figure.

For more information about Hub Console, refer to the following documentation:

Task	Topic(s)
Concepts	“Getting Started with the Hub Console” in the <i>Siperian Hub Administrator Guide</i>
Configuration	“Configuring Access to Hub Console Tools” in the <i>Siperian Hub Administrator Guide</i>

Master Reference Manager

Master Reference Manager (MRM) is the foundation product of Siperian Hub. Its purpose is to build an extensible and manageable system-of-record for all master data. It provides the platform to clean, match, consolidate, and manage master data across all data sources—internal and external—of an organization, and acts as a system-of-record for all downstream applications.

Hierarchy Manager

Siperian Hierarchy Manager (HM) is based on the foundation of Master Reference Manager. As the name implies, Hierarchy Manager allows users to manage hierarchy data that is associated with the records managed in MRM.

Hierarchy Manager provides a way to define hierarchical relationships and centrally manage data in a hierarchical manner. Many of the systems that are included in the master data management (MDM) landscape maintain the information about the relationships among the different data entities, as well as of the entities themselves. These disparate systems make it difficult to view and manage relationship data because each application has a different hierarchy, such as customer-to-account, sales-to-account or product-to-sales. Meanwhile, each data warehouse and data mart is designed to reflect relationships necessary for specific reporting purposes, such as sales by region by product over a specific period of time.

Hierarchy Manager includes two tools in the Hub Console:

Tool	Description
Hierarchies tool	Used by Siperian Hub administrators to set up the structures (entity types, hierarchies, relationships types, packages, and profiles) required to view and manipulate data relationships in Hierarchy Manager.
Hierarchy Manager tool	Used by data stewards to define and manage hierarchical relationships in their Hub Store.

The run-time component of Hierarchy Manager is bundled and deployed with the Hub Server application in the J2EE application server environment.

To manage the Hierarchy Manager, refer to the following documentation.

Task	Topic(s)
Configuration	“Configuring Hierarchies” in the <i>Siperian Hub Administrator Guide</i>
Usage	“Using Hierarchy Manager” in the <i>Siperian Hub Data Steward Guide</i>
Application Development	<i>Siperian Hub Services Integration Framework Guide</i> and the Siperian Hub Javadoc, particularly topics that describe Siperian Hub operations associated with Hierarchy Manager.

Activity Manager

Siperian Activity Manager (AM) evaluates data events, synchronizes master data, and delivers unified views of reference and activity data from disparate sources. Activity Manager:

- Provides the federated data access facilities that allow combining the master data managed in the Siperian Hub with the transactional and analytical data maintained in other systems.
- Monitors and evaluates any changes to the data—both within the Siperian Hub and in the various transactional applications—and then synchronize these changes across other systems or provides alerts or notifications—all based on configurable, user-defined rules and actions.

Activity Manager has an intuitive, powerful UI for defining, designing, delivering and managing unified views for downstream applications and systems. It includes two tools in the Hub Console:

Tool	Description
Remote Packages tool	Provides the ability to design and configure federated queries that are made available as remote packages.
Rules and Actions tool	Provides the ability to design and configure user-defined rules and actions, used to help evaluate data changes and synchronizes data with subscribing downstream applications and systems.

To manage the Activity Manager, refer to the *Siperian Activity Manager Modeler User Guide*.

Security Access Manager

Siperian Security Access Manager (SAM) is the part of Siperian Hub that provides comprehensive and highly-granular security mechanisms to ensure that only authenticated and authorized users have access to Siperian Hub data, resources, and functionality. Security Access Manager provide a mechanism for security decisions, and can integrate with security providers—third-party products that provide security services (authentication, authorization, and user profile services) for users accessing Siperian Hub.

Note: The way in which you configure and implement Siperian Hub security is governed by your organization’s particular security requirements, by the IT environment in which it is deployed, and by your organization’s security policies, procedures, and best practices.

To manage the Security Access Manager, refer to the following documentation.

Task	Topic(s)
Concepts	“About Setting Up Security” in “Setting Up Security” in the <i>Siperian Hub Administrator Guide</i>
Configuration	“Setting Up Security” in the <i>Siperian Hub Administrator Guide</i>
Application Development	“Using the Security Access Manager with the SIF SDK” in <i>Siperian Hub Services Integration Framework Guide</i> and the Siperian Hub Javadoc

Metadata Manager

The Metadata Manager (MET) is a tool in the Hub Console that allows administrators to manage metadata in their Siperian Hub implementation. Metadata describes the various schema design and configuration components—base objects and associated columns, cleanse functions, match rules, mappings, and so on—in the Hub Store.

Using the Metadata Manager, administrators can:

- Validate the metadata in a Siperian Hub repository and generate a report of *issues* (discrepancies or problems between the physical and logical schemas) that warrant attention.
- Compare repositories and generate change lists that describe the differences between them
- Copy design objects from one repository to another—such as promoting a design object from development to production, or exporting/importing design objects between Siperian Hub implementations. In a distributed development environment, developers can use the Metadata Manager tool to share and re-use design objects.
- Export the repository’s metadata to an XML file for subsequent import or archival purposes.
- Visualize the schema using a graphical model view of the repository.

For more information about the Metadata Manager, see the *Siperian Hub Metadata Manager Guide*.

Services Integration Framework

The Services Integration Framework (SIF) is the part of Siperian Hub that interfaces with external programs and applications. SIF enables external applications to implement the request/response interactions using any of the following architectural variations:

- Loosely coupled web services using the SOAP protocol.
- Tightly coupled Java remote procedure calls based on Enterprise JavaBeans (EJBs) or XML.
- Asynchronous Java Message Service (JMS)-based messages.

These capabilities enable Siperian Hub to support multiple modes of data access, expose numerous Siperian Hub data services via the SIF SDK, and produce events based on data changes in the Siperian Hub. This facilitates inbound and outbound integration with external applications and data sources, which can be used in both synchronous and asynchronous modes.

For more information about the Services Integration Framework, refer to the following documentation.

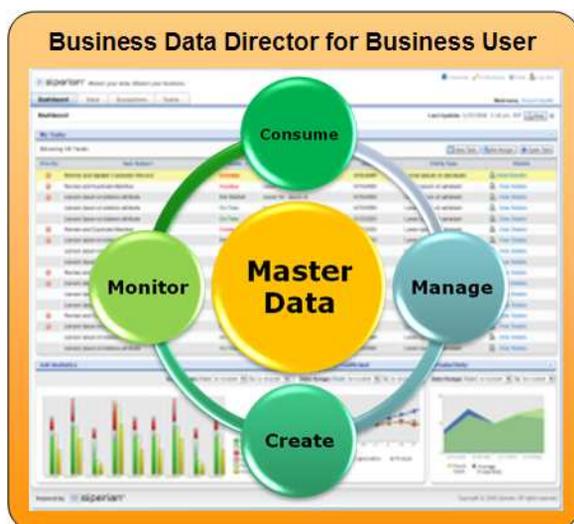
Task	Topic(s)
Concepts	“Introducing SIF SDK” in the <i>Siperian Hub Services Integration Framework Guide</i>
Configuration	“Setting up the SIF SDK” in the <i>Siperian Hub Services Integration Framework Guide</i> Part 5, “Configuring Application Access,” in the <i>Siperian Hub Administrator Guide</i>
Application Development	“Using the SIF SDK” in the <i>Siperian Hub Services Integration Framework Guide</i>
Reference	“About the Siperian Hub Operations” in the <i>Siperian Hub Services Integration Framework Guide</i> Siperian Hub Javadoc

Business Data Director

The Business Data Director (BDD) is a data governance application for Siperian Hub that enables business users to effectively create, manage, consume, and monitor master data. Business Data Director is web-based, task-oriented, workflow-driven, highly customizable, and highly configurable, providing a web-based configuration wizard that creates an easy-to-use interface based on your organization's data model.

Integrated task management ensures that all data changes are automatically routed to the appropriate personnel for approval prior to impacting to the 'best version of the truth.' As tasks are routed, the Business Data Director Dashboard provides business users with a view of assigned tasks, while also providing a graphical view into key metrics such as productivity and data quality trending.

In addition, Business Data Director leverages Siperian's Security Access Manager (SAM) module, providing a comprehensive and flexible security framework - enabling both attribute and data level security. With this, customers can strike that elusive balance between open and secure by strengthening policy compliance and ensuring access to critical information.



Business Data Director enables data stewards and other business users to:

- **Create Master Data.** Working individually or collaboratively across lines of business, users can add new entities and records to the Hub Store. Offering capabilities such as inline data cleansing and duplicate record identification and resolution during data entry, Business Data Director enables users to proactively validate, augment, and enrich their master data.
- **Manage Master Data.** Users can approve and manage updates to master data, manage hierarchies via drag and drop, resolve potential matches and merge duplicates, and create and assign tasks to other users.
- **Consume Master Data.** Users can search for all master data from a central location, and then view master data details and hierarchies. Users can also embed UI components into business applications.
- **Monitor Master Data.** Users can track the lineage and history of master data, audit their master data for compliance, and use a customizable dashboard that shows them the most relevant information.

With the Business Data Director, companies can reduce cost of quality by proactively managing data, improve productivity by finding accurate information faster, enable compliance by providing a complete, consistent view of data and lineage, and increase revenue by acting on master data relationship insights.

Key Concepts

This chapter describes the concepts that users need to understand in order to work effectively with the Siperian Hub.

Chapter Contents

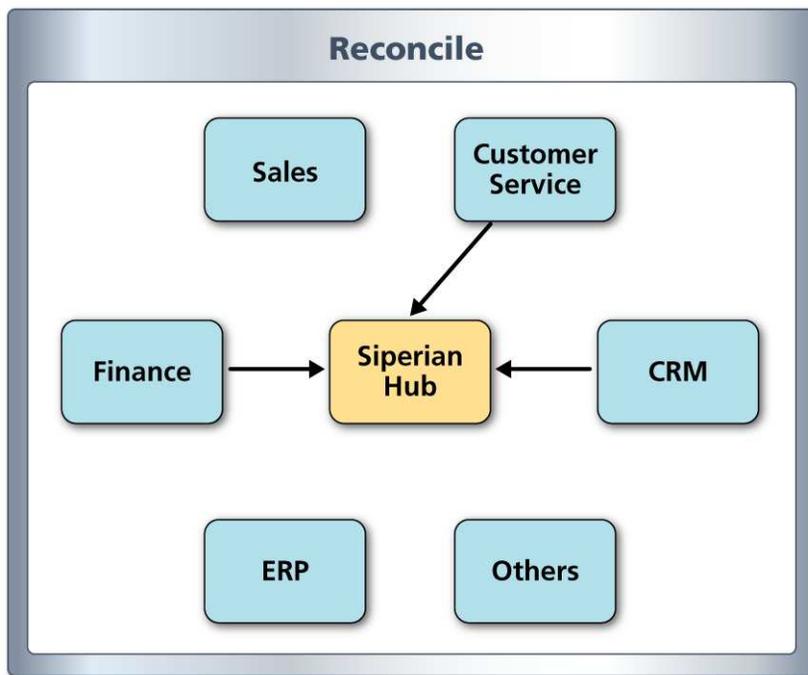
- [Inbound and Outbound Data Flows](#)
- [Batch and Real-Time Processing](#)
- [Batch Processing](#)
- [Real-Time Processing](#)
- [Databases in the Hub Store](#)
- [Content Metadata](#)
- [Workflow Integration and State Management](#)
- [Hierarchy Management](#)
- [Activity Management](#)

Inbound and Outbound Data Flows

This section describes the main inbound and outbound data flows for Siperian Hub. For information about the processes that make up these data flows, see [“Batch Processing” on page 29](#).

Main Inbound Data Flow (Reconciliation)

The main inbound flow into Siperian Hub is called *reconciliation*.



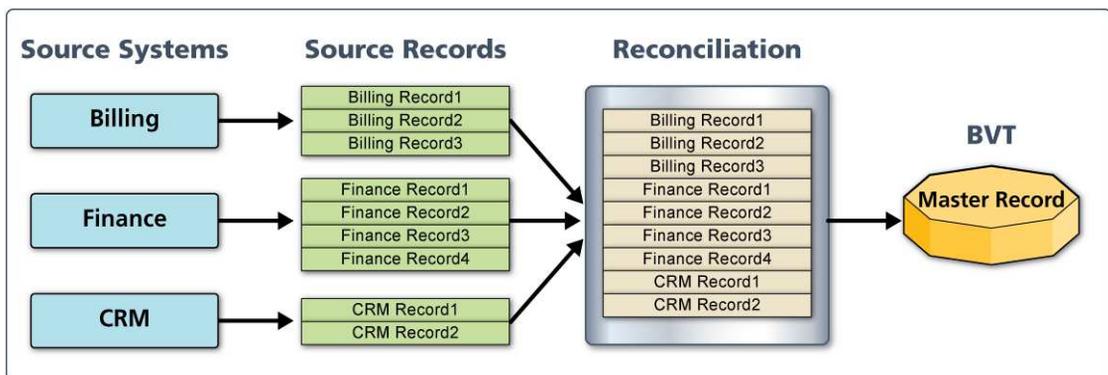
In Siperian Hub, business entities—such as customers, accounts, products, or employees—are represented in tables called *base objects*. For a given base object:

- Siperian Hub obtains data from one or more *source systems*—an operational system or third-party application that provides data to Siperian Hub for cleansing, matching, consolidating, and maintenance.

Reconciliation can involve cleansing the data beforehand to optimize the process of matching and consolidating records. Cleansing is the process by which data is standardized by validating, correcting, completing, or enriching it.

- An individual entity (such as a specific customer or account) can be represented by multiple records (“multiple versions of the truth”) in the base object.
- Siperian Hub then reconciles “multiple versions of the truth” to arrive at the master record—the best version of the truth—for each individual entity. Consolidation is the process of merging duplicate records to create a *consolidated record* that contains the most reliable cell values from the source records.

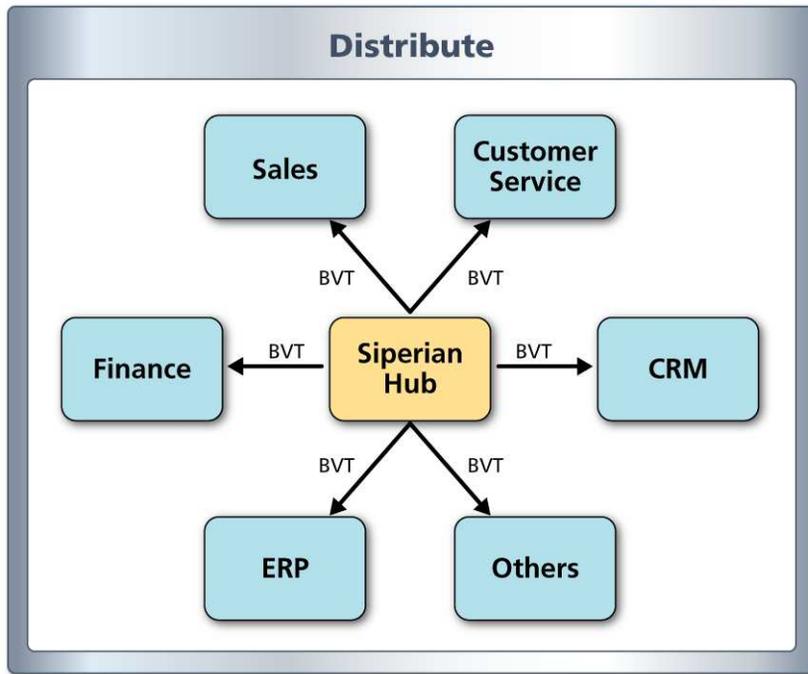
For example, suppose the billing, finance, and customer relationship management applications all have different billing addresses for a given customer. Siperian Hub can be configured to determine which data represents the best version of the truth based on the relative reliability of column data from different source systems based on such factors as the age of the data (the customer’s most recent purchase).



Data in the master record might derive from a single record (such as the most recent billing address from the billing system), or it might represent a composite of data from different records.

Main Outbound Data Flow (Distribution)

The main outbound flow out of Siperian Hub is called *distribution*. Once the master record is established for a given entity, Siperian Hub can then (optionally) distribute the master record data to other applications or databases.



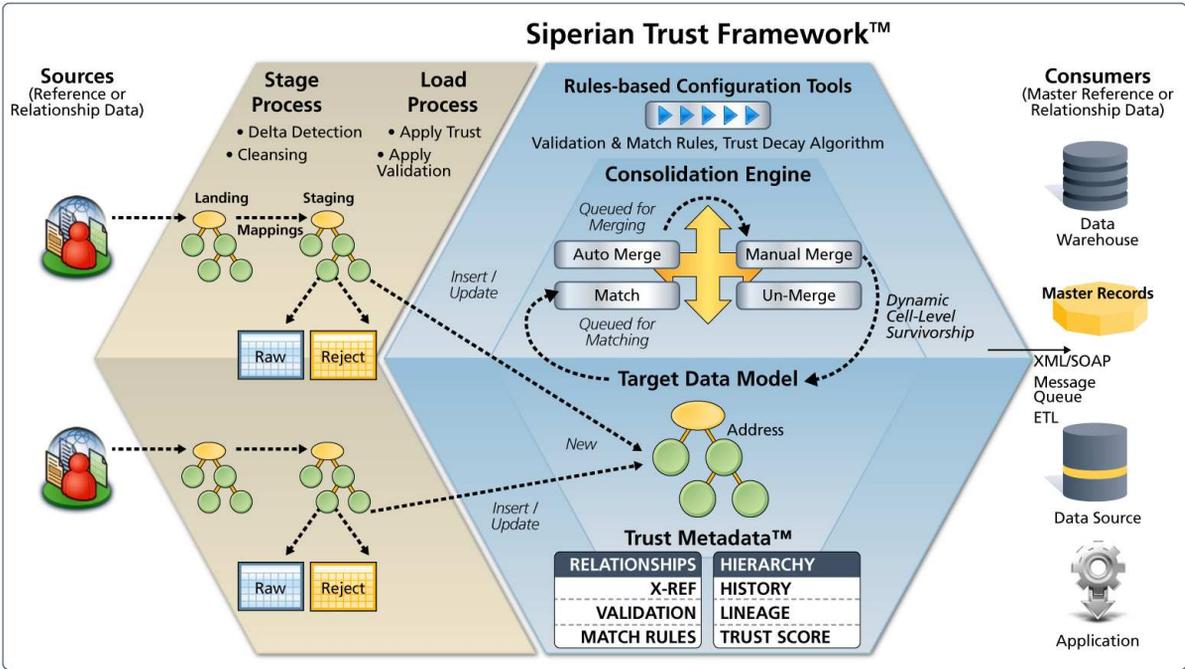
For example, if an organization's billing address has changed in Siperian Hub, then Siperian Hub can notify other systems in the organization (via JMS messaging) about the updated information so that master data is synchronized across the enterprise.

Batch and Real-Time Processing

Siperian Hub has a well-defined data management flow that proceeds through distinct processes in order for the data to get reconciled and distributed. Data can be processed by Siperian Hub into two different ways: batch processing and real-time processing. Many Siperian Hub implementations use a combination of both batch and real-time processing as applicable to the organization's requirements.

Batch Process Flow

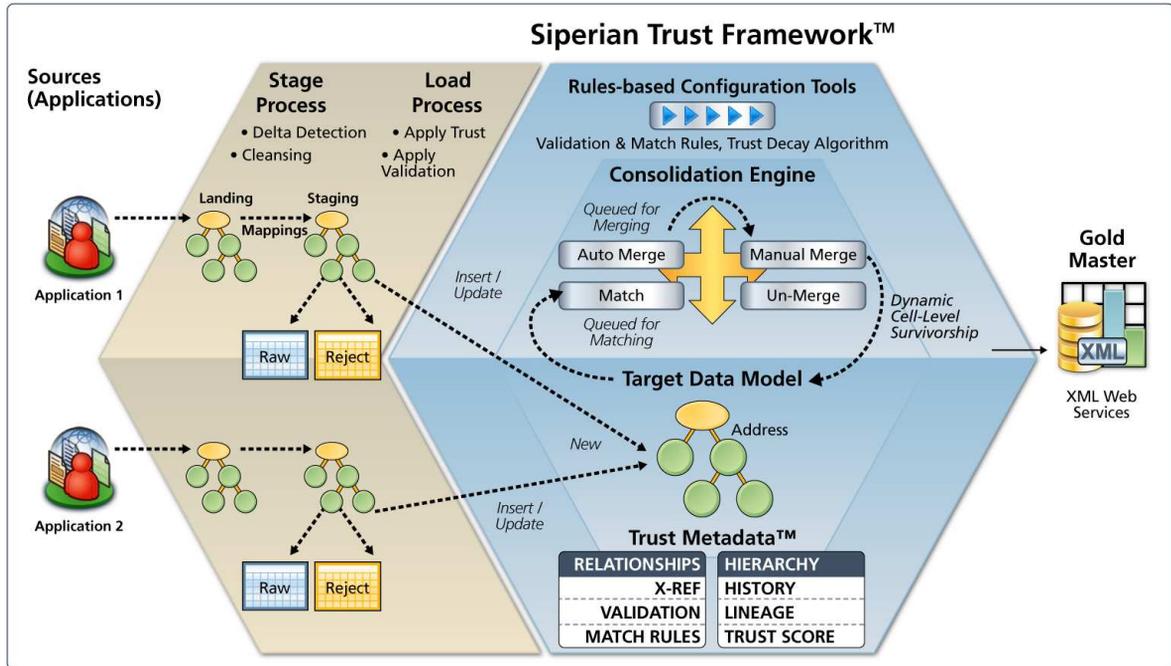
The following figure shows the overall batch process flow for processing data in Siperian Hub.



For more information, see [“Batch Processing”](#) on page 29.

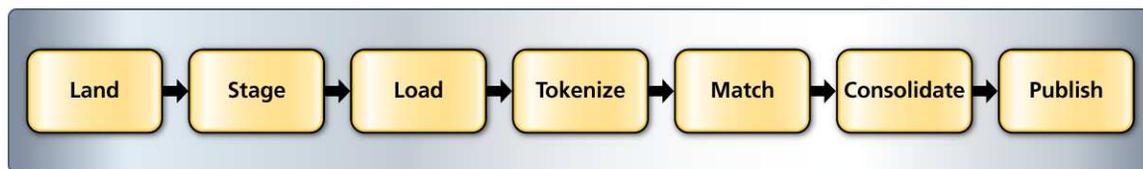
Real-time Process Flow

The following figure shows the overall real-time process flow for processing data in Siperian Hub.



For more information, see [“Real-Time Processing”](#) on page 37.

Batch Processing



For batch processing, data is loaded from source systems and processed in Siperian Hub via a series of processes.

Process	Description
Land	Transfers data from a source system (external to Siperian Hub) to landing tables in the Hub Store. Part of the reconciliation process described in “Main Inbound Data Flow (Reconciliation)” on page 24.
Stage	Retrieves data from the landing table, cleanses it (if applicable), and copies it into a staging table in the Hub Store. Part of the reconciliation process.
Load	Loads data from the staging table into the corresponding Hub Store table (base object or dependent object). Part of the reconciliation process.
Tokenize	Generates match tokens in a match key table that are used subsequently by the match process to identify candidate base object records for matching.
Match	Compares records for points of similarity (based on match rules), determines whether records are duplicates, and flags duplicate records for consolidation. Part of the reconciliation process.
Consolidate	Merges data in duplicate records to create a <i>consolidated record</i> that contains the most reliable cell values from the source records. Part of the reconciliation process.
Publish	Publishes the BVT to other systems or processes via outbound JMS message queues. Part of the distribution process described in “Main Outbound Data Flow (Distribution)” on page 26.

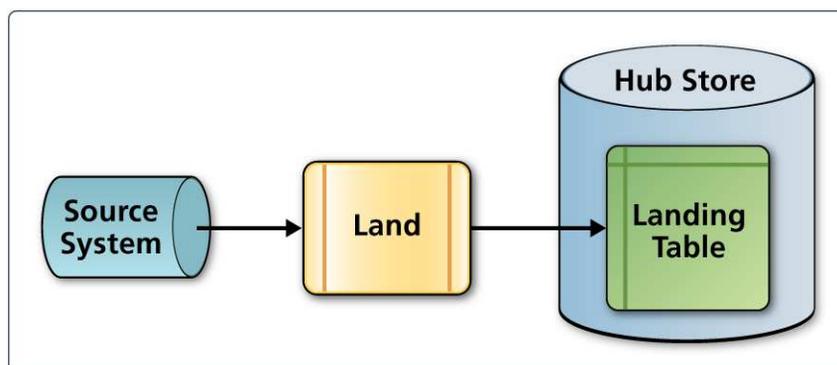
Siperian Hub batch processes are implemented as database stored procedures that can be invoked from the Hub Console or through custom scripts using third-party job management tools.

In Siperian Hub implementations, batch processing is used as appropriate. For example, batch processing is often used for the *initial data load* (the first time that business data is loaded into the Hub Store), as it can be the most efficient way to load a large number of records into Siperian Hub. Batch processing is also used when it is the only way—or the most efficient way—to get data from a particular source system.

For more information about batch processes, see the following topics:

Task	Topic(s)
Concepts	“Siperian Hub Processes” in the <i>Siperian Hub Administrator Guide</i>
Configuration	<ul style="list-style-type: none"> • “Configuring the Land Process” in the <i>Siperian Hub Administrator Guide</i> • “Configuring the Stage Process” and “Configuring the Cleanse Process” in the <i>Siperian Hub Administrator Guide</i> • “Configuring the Load Process” in the <i>Siperian Hub Administrator Guide</i> • “Configuring the Match Process” in the <i>Siperian Hub Administrator Guide</i> • “Configuring the Consolidate Process” in the <i>Siperian Hub Administrator Guide</i> • “Configuring the Publish Process” in the <i>Siperian Hub Administrator Guide</i>
Execution	<ul style="list-style-type: none"> • “Using Batch Jobs” in the <i>Siperian Hub Administrator Guide</i> • “Scheduling Batch Jobs and Batch Groups” in the <i>Siperian Hub Administrator Guide</i> • <i>Siperian Hub Data Steward Guide</i> • Documentation for the ETL tool(s) or external process(es) used to land data in landing tables
Application Development	<i>Siperian Hub Services Integration Framework Guide</i> and the Siperian Hub Javadoc

Land Process



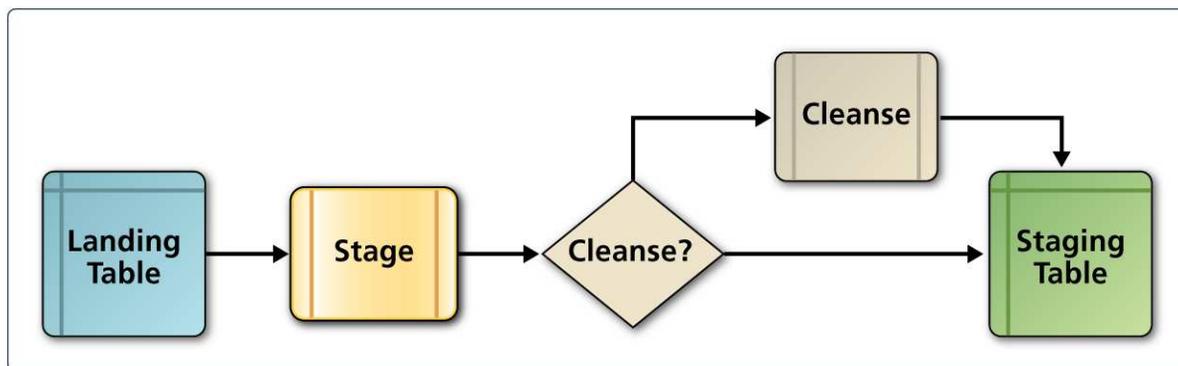
The land process transfers data from a source system to landing tables in the Hub Store. A *landing table* provides intermediate storage in the flow of data from source systems into Siperian Hub. In effect, landing tables are “where data lands” from contributing source systems.

Landing tables are populated during the land process in either of two ways:

Mode	Description
batch processing	A third-party ETL (Extract-Transform-Load) tool or other external process writes the data into one or more landing tables. Such tools or processes are not part of the Siperian Hub suite of products.
on-line, real-time processing	An external application populates landing tables in the Hub Store. This application is not part of the Siperian Hub suite of products.

The land process is *external* to Siperian Hub and is executed via an external batch process (such as a third-party ETL—Extract-Transform-Load—tool), or in on-line, real-time mode (in which an external application directly populates landing tables in the Hub Store). Subsequent processes for managing data are internal to Siperian Hub.

Stage Process



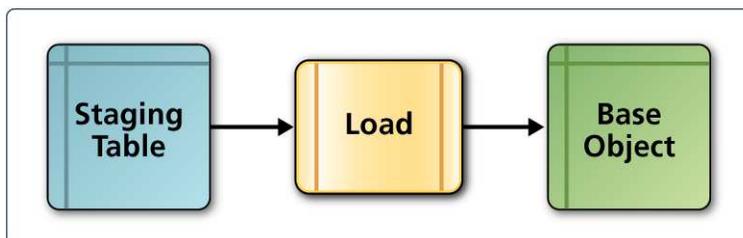
The stage process reads the data from the landing table, cleanses the data if applicable, and moves the cleansed data into a staging table in the Hub Store. The *staging table* provides temporary, intermediate storage in the flow of data from landing tables into base objects.

Mappings facilitate the transfer and cleansing of data between landing and staging tables during the stage process. A mapping defines:

- which landing table column is used to populate a column in the staging table
- what standardization and verification (cleansing) must be done, if any, before the staging table is populated.

Siperian Hub standardizes and verifies data using *cleanse functions*. Each cleanse function provides access to specialized cleansing functionality, such as address verification, address decomposition, gender determination, title/upper/lower-casing, white space compression, and so forth. The output of the cleanse function becomes the input to the target column in the staging table.

Load Process



The load process loads data from the staging table into the corresponding Hub Store table, called a *base object* (or a *dependent object*).

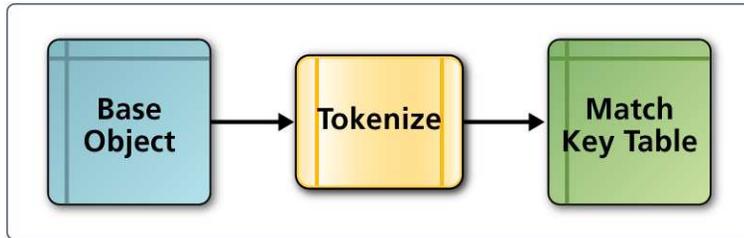
If a column in a base object derives its data from multiple source systems, Siperian Hub uses *trust* to help with comparing the relative reliability of column data from different source systems. For example, the Orders system might be a more reliable source of billing addresses than the Sales system.

Trust provides a mechanism for measuring the confidence factor associated with each cell based on its source system, change history, and other business rules. Trust takes into account the age of data, how much its reliability has decayed over time, and the validity of the data. Trust is used to determine survivorship (when two records are consolidated) and whether updates from a source system are sufficiently reliable to update the master record.

Trust is often used in conjunction with validation rules, which tell Siperian Hub the condition under which a data value is not valid. When data meets the criterion specified by the validation rule, then the trust value for that data is downgraded by the percentage specified in the validation rule. For example:

```
Downgrade trust on First_Name by 50% if Length < 3
```

Tokenize Process

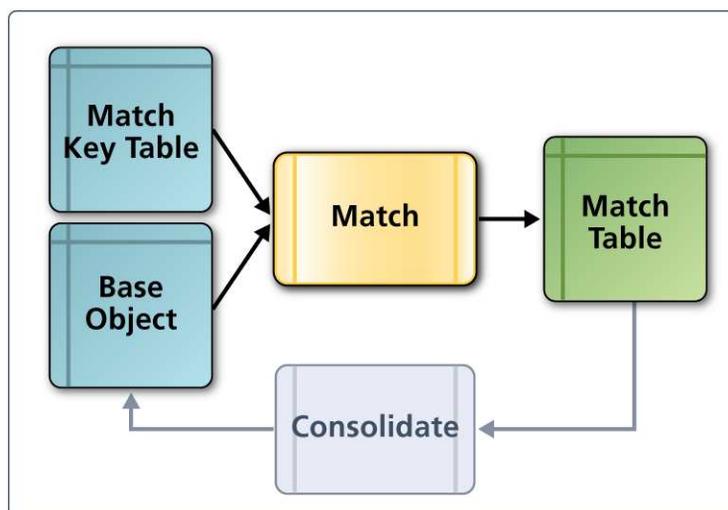


The *tokenize process* generates match tokens that are used subsequently by the match process to identify candidate base object records for matching. *Match tokens* are strings that represent both encoded (match key) and unencoded (raw) values in the match columns of the base object. *Match keys* are fixed-length, compressed, and encoded values, built from a combination of the words and numbers in a name or address, such that relevant variations have the same match key value.

The generated match tokens are stored in a *match key table* associated with the base object. For each record in the base object, the tokenize process stores one or more records containing generated match tokens in the match key table. The match process depends on current data in the match key table, and will run the tokenize process automatically if match tokens have not been generated for any of the records in the base object. The tokenize process can be run before the match process—automatically, at the end of the load process, or manually, as a batch job or stored procedure.

The Hub Console allows users to investigate the distribution of match keys in the match key table. Users can identify potential *hot spots* in their data—high concentrations of match keys that could result in *overmatching*—where the match process generates too many matches, including matches that are not relevant.

Match Process

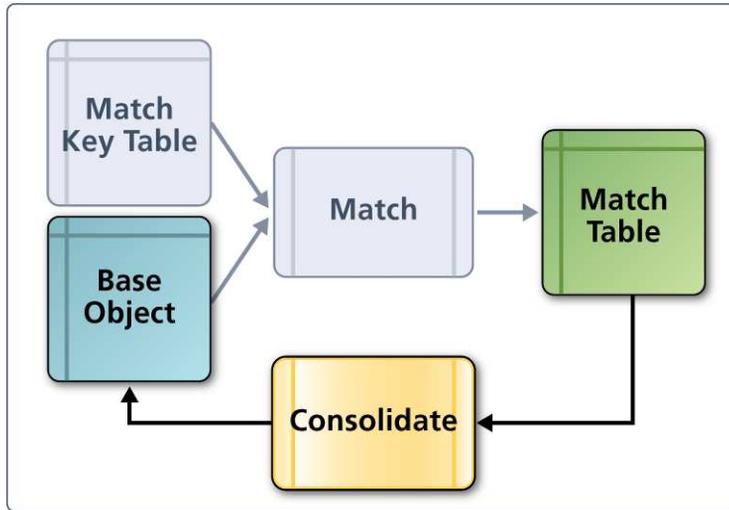


The match process identifies data that conforms to the match rules that you have defined. These rules define duplicate data for Siperian Hub to consolidate. *Matching* is the process of comparing two records for points of similarity. If sufficient points of similarity are found to indicate that the two records are probably duplicates of each other, then Siperian Hub flags those records for consolidation.

In a base object, the columns to be used for comparison purposes are called *match columns*. Each match column is based on one or more columns from the base object. Match columns are combined into *match rules* to determine the conditions under which two records are considered to be similar enough to consolidate. Each match rule tells Siperian Hub the combination of match columns it needs to examine for points of similarity. When Siperian Hub finds two records that satisfy a match rule, it records the primary keys of the records, as well as the match rule identifier. The records are flagged for either automatic or manual consolidation according to the category of the match rule.

External match is used to match new data with existing data in a base object, test for matches, and inspect the results—all without actually loading the data into the base object. External matching is used to pretest data, test match rules, and inspect the results before running the actual match process on the data.

Consolidate Process



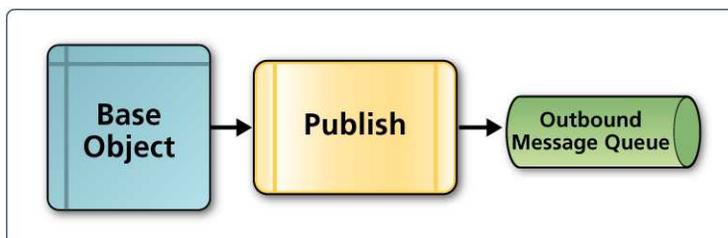
After duplicate records have been identified in the match process, the consolidate process merges duplicate records into a single record.

System	Master ID	First Name	MI	Last Name	Address	City	State	Zip
Siperian	M-0001	Abel	Noel	Willan	161 Washington Ave	Buffalo	NY	14263
Sales Force	SFA_ID	Abel		Willan	161 Washington Ave	Buffalo	NY	14263
Accounts	Cust_ID	Abel	Noel	Willan	161 Washington Ave	Buffalo	NY	14263
Marketing	Target_ID	Abel	N	Willan	Elm and Carlton Streets	Buffalo	NY	14263

The goal in Siperian Hub is to identify and eliminate all duplicate data and to merge them together into a single, consolidated master record containing the most reliable cell

values from the source records. For more information about the consolidate process, see “Configuring the Consolidate Process” in the *Siperian Hub Administrator Guide*.

Publish Process



The publish process can be configured to publish the BVT to an outbound JMS message queue. Other external systems, processes, or applications that listen on the message queue can retrieve the message and process it accordingly. For more information about the publish process, see “Configuring the Publish Process” in the *Siperian Hub Administrator Guide*.

Real-Time Processing

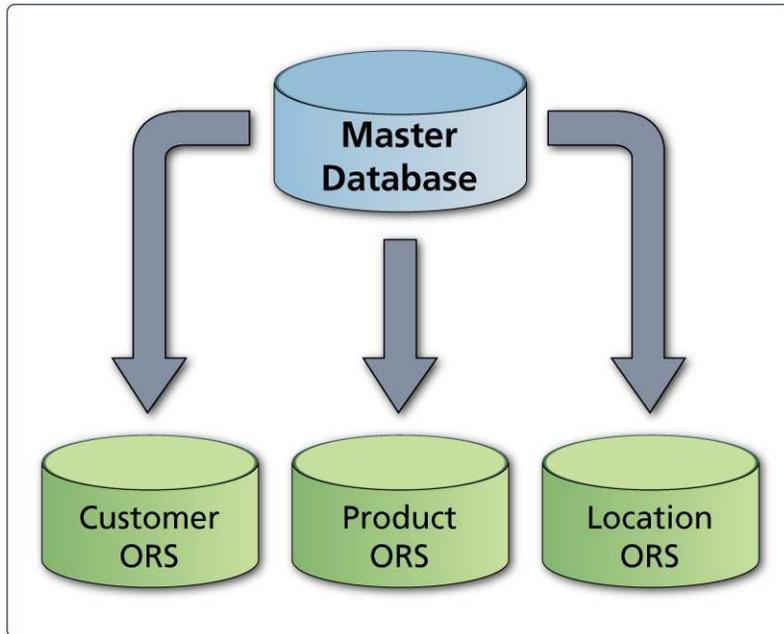
For real-time processing, applications that are external to Siperian Hub invoke Siperian Hub operations via the Services Integration Framework (SIF) interface. SIF provides APIs for various Siperian Hub services, such as reading, cleansing, matching, inserting, and updating records.

In Siperian Hub implementations, real-time processing is used as appropriate. For example, real-time processing can be used to update data in the Hub Store whenever a record is added, updated, or deleted in a source system. Real-time processing can also be used to handle *incremental data loads* (data loads that occur after the initial data load) into the Hub Store.

For more information about SIF, see the *Siperian Hub Services Integration Framework Guide* and the Siperian Hub Javadoc. Siperian Hub can generate events to notify external applications when specific data changes occur in the Hub Store.

Databases in the Hub Store

The Hub Store is a collection of databases that contain configuration settings and data processing rules.



The *Master Database* is a database in the Hub Store that contains the Siperian Hub environment configuration settings—user accounts, security configuration, ORS registry, message queue settings, and so on. A given Siperian Hub environment can have only one Master Database.

An *Operational Record Store (ORS)* is a database in the Hub Store that contains the master data, content metadata, rules for processing the master data, the rules for managing the set of master data objects, along with the processing rules and auxiliary logic used by the Siperian Hub in defining the best version of the truth (BVT). A Siperian Hub configuration can have one or more ORS databases.

Content Metadata

For each base object in the schema, Siperian Hub automatically maintains support tables containing *content metadata* about data that has been loaded into the Hub Store. For more information about content metadata and support tables, see “Building the Schema” in the *Siperian Hub Administrator Guide*.

Base Objects

A *base object* (sometimes abbreviated as *BO*) is a table in the Hub Store that is used to describe central business entities, such as customers, accounts, products, employees, and so on. The base object is the end-point for consolidating data from multiple source systems. In a Siperian Hub implementation, the *schema* (or data model) for an organization typically includes a collection of base objects.

The goal in Siperian Hub is to create the *master record* for each instance of each unique entity within a base object. As mentioned in [“Master Data and Master Data Management” on page 2](#), the master record is said to contain the *best version of the truth* (abbreviated as *BVT*), which is a record that has been consolidated with the best, most-trustworthy cell values from the source records. For example, for a Customer base object, you want to end up with a master record for each individual customer. The master record in the base object contains the best version of the truth for that customer.

Cross-Reference (XREF) Tables

Cross-reference tables, sometimes referred to as XREF tables, are used for tracking the lineage of data—which systems, and which records from those systems, contributed to consolidated records.

For each source system record, Siperian Hub maintains a cross-reference record that contains an identifier for the system that provided the record, the primary key value of that record in the source system, and the most recent cell values provided by that system. If the same column (for example, phone number) is provided by multiple source systems, the XREF table contains the value from every source system.

Each base object record will have one or more cross-reference records. Cross-reference tables are used for merge and unmerge operations, as well as delete management (removing records that were contributed by a particular source system).

History Tables

History tables are used for tracking this history of changes to a base object and its lineage back to the source system. Siperian manages several different history tables—including base object and cross-reference history tables—to provide detailed change-tracking options, including merge and unmerge history, history of the pre-cleansed data, history of the base object, and history of the cross-reference.

Workflow Integration and State Management

Siperian Hub supports workflow tools by storing pre-defined system states—ACTIVE, PENDING, and DELETED—for base object and XREF records. By enabling state management on your data, Siperian Hub allows integration with workflow integration processes and tools, supports a “change approval” process to ensure that only approved records contribute to the best version of the truth, and tracks intermediate stages of the process (pending records). For more information, see “State Management” in the *Siperian Hub Administrator Guide* and the *Siperian Hub Data Steward Guide*.

Hierarchy Management

As described in [“Hierarchy Manager” on page 16](#), the Hierarchy Manager (HM) allows users to manage hierarchy data that is associated with the records managed in MRM. For more information, see [“Configuring Hierarchies”](#) in the *Siberian Hub Administrator Guide* and [“Using Hierarchy Manager”](#) in the *Siberian Hub Data Steward Guide*.

Relationships

In Hierarchy Manager, a *relationship* describes the affiliation between two specific entities. Hierarchy Manager relationships are defined by specifying the relationship type, hierarchy type, attributes of the relationship, and dates for when the relationship is active. Information about a Hierarchy Manager entity is stored in a *relationship base object*. A *relationship type* describes classes of relationships. A relationship type defines the types of entities that a relationship of this type can include, the direction of the relationship (if any), and how the relationship is displayed in the Hub Console.

Hierarchies

A *hierarchy* is a set of relationship types. These relationship types are not ranked, nor are they necessarily related to each other. They are merely relationship types that are grouped together for ease of classification and identification. The same relationship type can be associated with multiple hierarchies. A *hierarchy type* is a logical classification of hierarchies.

Entities

In Hierarchy Manager, an *entity* is any object, person, place, organization, or other thing that has meaning and can be acted upon in your database. Examples include a specific person’s name, a specific checking account number, a specific company, a specific address, and so on. Information about a Hierarchy Manager entity is stored in an *entity base object*, which you create and configure in the Hub Console. An *entity type* is a logical classification of one or more entities. Examples include doctors, checking accounts, banks, and so on. All entities with the same entity type are stored in the same entity object.

Activity Management

As described in “[Activity Manager](#)” on page 17, Siperian Activity Manager (AM) evaluates data events, synchronizes master data, and delivers unified views of reference and activity data from disparate sources. To manage the Activity Manager, refer to the *Siperian Activity Manager Modeler User Guide*.

Rules and Actions

Activity Manager provides a sophisticated rules engine that, in conjunction with the Federated Query Engine, can monitor and evaluate changes to the master reference data within the Siperian Hub or third-party applications. As events occur to the master reference data in the Hub Store or transactional data within a third-party application, triggers execute the rules that have been defined within Activity Manager. In turn, these rules can be configured to invoke actions such as invoking a web service.

The Rules and Actions tool in the Hub Console allow users to design, manage, and deploy rules and actions. A *rule* consists of zero or more conditions and is based on package data and event metadata which, if successfully evaluate to being true, will request the execution of one or more actions. An *action* is something that will be performed as a consequence of successfully evaluating a rule to be true (such as a Web Service call, logging, and so on). Rules are activated by a *trigger*, which is an event or input that causes the rule engine to start rules evaluation.

Remote Packages

Using the Remote Packages Console and Activity Manager Modeler, implementers can design unified views of the master data. The Activity Manager Modeler is an Eclipse application that allows implementers to edit existing data models and create *remote packages*. A remote package combines data from diverse external systems, using the Activity Manager federated query engine to create a single view of the data. Implementers can quickly join several views of data from third party applications into a single view stored in a remote package. The Activity Manager Modeler transforms the separate views of data into a single SQL statement to be consumed by the remote package.

Topics for Siperian Hub Users

This chapter describes types of Siperian Hub users and points to resources of interest to each type of user.

Chapter Contents

- [Administrators](#)
- [Developers](#)
- [Data Stewards](#)

Administrators

This section describes activities and resources for Siperian Hub administrators.

About Siperian Hub Administrators

Administrators have primary responsibility for the set up and configuration of the Siperian Hub system, including:

- installing the Siperian Hub software
- setting up the database and Hub Store
- building the data model and other objects in the Hub Store
- configuring and executing Siperian Hub data management processes
- configuring security
- configuring external application access to Siperian Hub operations and resources
- monitoring ongoing operations

Administrators access Siperian Hub through the Hub Console, which comprises a set of tools for managing a Siperian Hub implementation.

Documentation Resources for Siperian Hub Administrators

Task	Topic(s)
Concepts	<i>Siperian Hub Overview</i>
Installation	<i>Siperian Hub Installation Guide</i> for your platform <i>Siperian Hub Cleanse Adapter Guide</i> <i>Siperian Hub Release Notes</i> <i>What's New in Siperian Hub</i>
Administration	<i>Siperian Hub Administrator Guide</i> <i>Siperian Hub Metadata Manager Guide</i>

Developers

This section describes activities and resources for Siperian Hub developers.

About Siperian Hub Developers

Developers have primary responsibility for designing, developing, testing, and deploying external applications that integrate with Siperian Hub.

Documentation Resources for Siperian Hub Developers

Task	Topic(s)
Concepts	<i>Siperian Hub Overview</i> , especially “Services Integration Framework” on page 20 .
Configuration	Part 5, “Configuring Application Access,” in the <i>Siperian Hub Administrator Guide</i>
Application Development	<i>Siperian Hub Services Integration Framework Guide</i> <i>Siperian Hub Resource Kit User Guide</i>
Reference	Siperian Hub Javadoc

Data Stewards

This section describes activities and resources for data stewards using Siperian Hub tools.

About Siperian Hub Data Stewards

Data stewards have primary responsibility for data quality. Data stewards can access Siperian Hub in two ways:

- **Business Data Director**, described in [“Business Data Director” on page 21](#)
- **Hub Console**, which includes the following tools:

Tool	Description
Merge Manager	Used to review and take action on the records that are queued for manual merging, as well as monitor the records that are queued for automerge. Data stewards can view newly-loaded base object records that have been matched against other records in the base object and, based on this view, can <ul style="list-style-type: none">• combine duplicate records together to create consolidated records• designate records that are not duplicates as unique records
Data Manager	Used to review the results of all merges and links—including automatic merges and links—and to correct data if necessary. Data stewards can view the data lineage for each base object record, unmerge previously-consolidated records, and view different types of history on each consolidated record.
Hierarchy Manager	Used to define and manage hierarchical relationships in the Hub Store.

Documentation Resources for Siperian Hub Data Stewards

Task	Topic(s)
Concepts	<i>Siperian Hub Overview</i>
Usage	<i>Siperian Hub Data Steward Guide</i>

A

actions 42
Activity Manager (AM) 17
administrators 44
audience vii

B

base objects 24, 39
batch processing
 about batch processing 29
 consolidate process 36
 land process 31
 load process 33
 match process 35
 overall flow 27
 publish process 37
 stage process 32
 tokenize process 34
best version of the truth (BVT) 25
Business Data Director 21

C

cleanse functions 32
Cleanse Match Server 13
consolidate process 36
consolidated record 25

content metadata 39
cross-reference tables 39
customer support xi

D

data model 39
data stewards 46
database administrators 44
dependent objects 33
developers 45
distribution 26
documentation
 audience vii
 list of documents viii
 organization viii

E

entities 41
ETL tools 31
external match 35
extraction-transformation-load tools 31

H

hierarchies 41
Hierarchy Manager (HM) 16
history tables 40

hotspots 34
Hub Console 14
Hub Server 12
Hub Store 11

I

incremental data loads 37
initial data loads 30
introduction 1

J

JMS message queues 37

L

land process 31
landing tables 31
load process 33

M

mappings 32
master data 2
Master Data Management (MDM) 2
Master Database 38
master records 25
Master Reference Manager (MRM) 15
match columns 35
match key tables 34
match keys 34
match process 35
match rules 35
match tokens 34

merging duplicate records 36
message queues 37
Metadata Manager (MET) 19

O

Operational Record Store (ORS) 38
overmatching 34

P

preface vii
publish process 37

R

real-time processing
 about real-time processing 37
 overall flow 28
reconciliation 24
relationships 41
remote packages 42
rules 42

S

schema 39
Security Access Manager (SAM) 18
Services Integration Framework (SIF) 20, 37
Siperian Hub
 about Siperian Hub 5
 architecture 10
 components of 10
 core capabilities 6
Siperian, contacting xi

source systems 24
stage process 32
staging tables 32
state management 40
support xi
system administrators 44

T

technical support xi
tokenize process 34
training x
trust 33

V

validation rules 33

W

workflow integration 40

X

XREF tables 39

