

MRS: Huawei's Big Data Platform

Huawei DataArts Studio

Prepared by : Hagar Muhamed

Lecture 10





مجموعة جامعة
الجميع الذكية
للتعليم القابضة



MRS: Huawei's Big Data Platform

Huawei DataArts Studio

Lecture 10

Prepared by
Dr. Hagar Muhamed

www.midocean.ae

Overview of MRS

MRS Components

Huawei DataArts Studio

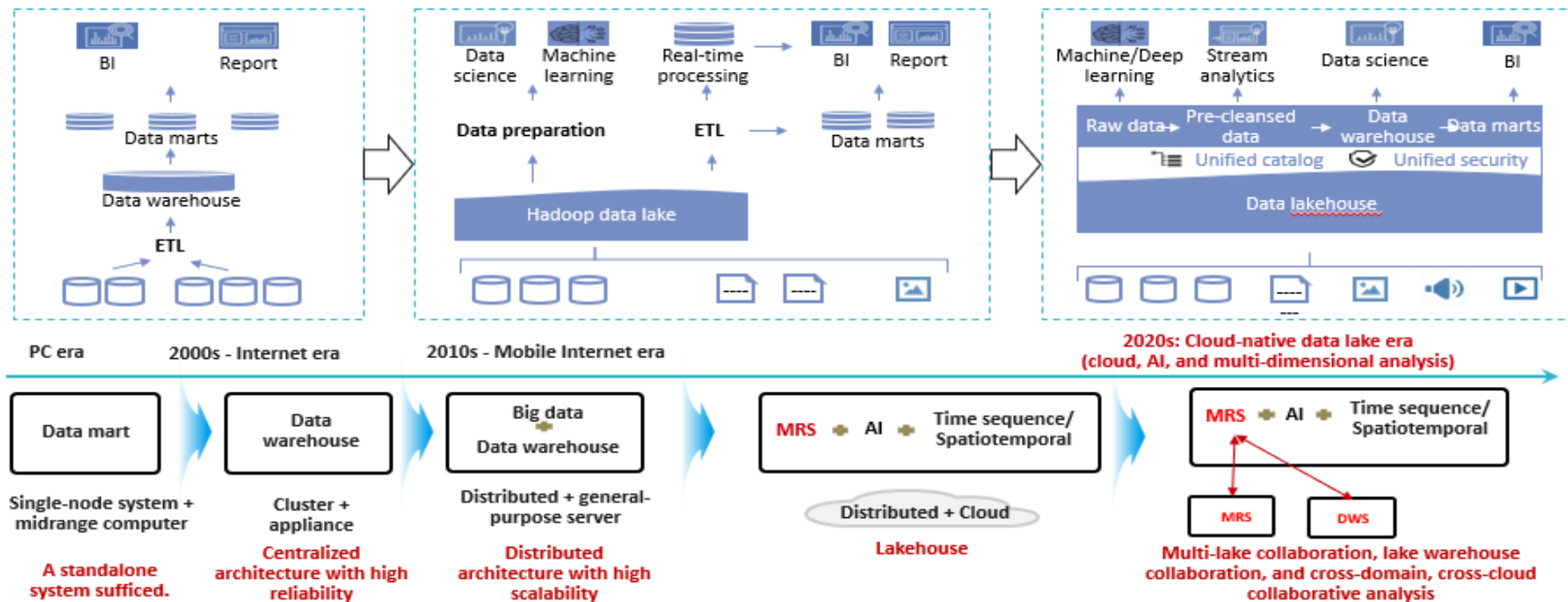
Chapter

10

Content



Trends in the Evolution of Big Data Technology



Huawei Cloud Services

Huawei Cloud is Huawei's signature cloud service brand. It is a culmination of Huawei's 30-plus years of expertise in ICT infrastructure products and solutions. Huawei Cloud is committed to providing stable, secure, and reliable cloud services that help organizations of all sizes grow in an intelligent world. To complement an already impressive list of offerings, Huawei Cloud is pursuing a vision of inclusive AI, a vision of AI that is affordable, effective, and reliable for everyone. As a foundation, Huawei Cloud provides a powerful computing platform and an easy-to-use development platform for Huawei's full-stack all-scenario AI strategy.

Huawei Cloud MRS

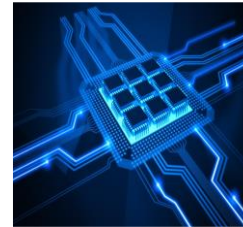
MapReduce Service (MRS) is used to deploy and manage Hadoop systems on Huawei Cloud. MRS provides enterprise-level big data clusters on the cloud. Tenants have full control over clusters and can easily run big data components such as Hadoop, Spark, HBase, Kafka, and Storm. MRS is fully compatible with open-source APIs, and incorporates the advantages of Huawei Cloud computing and storage and big data industry experience to provide customers with a full-stack big data platform featuring high performance, low cost, flexibility, and ease-of-use. In addition, the platform can be customized based on service requirements to help enterprises quickly build a massive data processing system and discover new value points and business opportunities by analyzing and mining massive amounts of data in real time or at a later time.

MRS Highlights



Decoupled storage and compute

- A unified data lake eliminates data silos; a single copy of data is enough, with no need for data transfer; multiple computing engines, making flexible allocation and on-demand scaling of storage and computing resources possible; 30% more cost-effective than the industry average



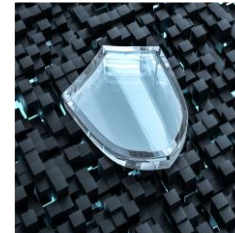
First-class performance and experience

- Full-stack performance acceleration and millisecond-level response to millions of metadata records by means of four-level vertical optimization on hardware, data organization, computing engine, and AI intelligent optimization, providing users with the ultimate performance experience



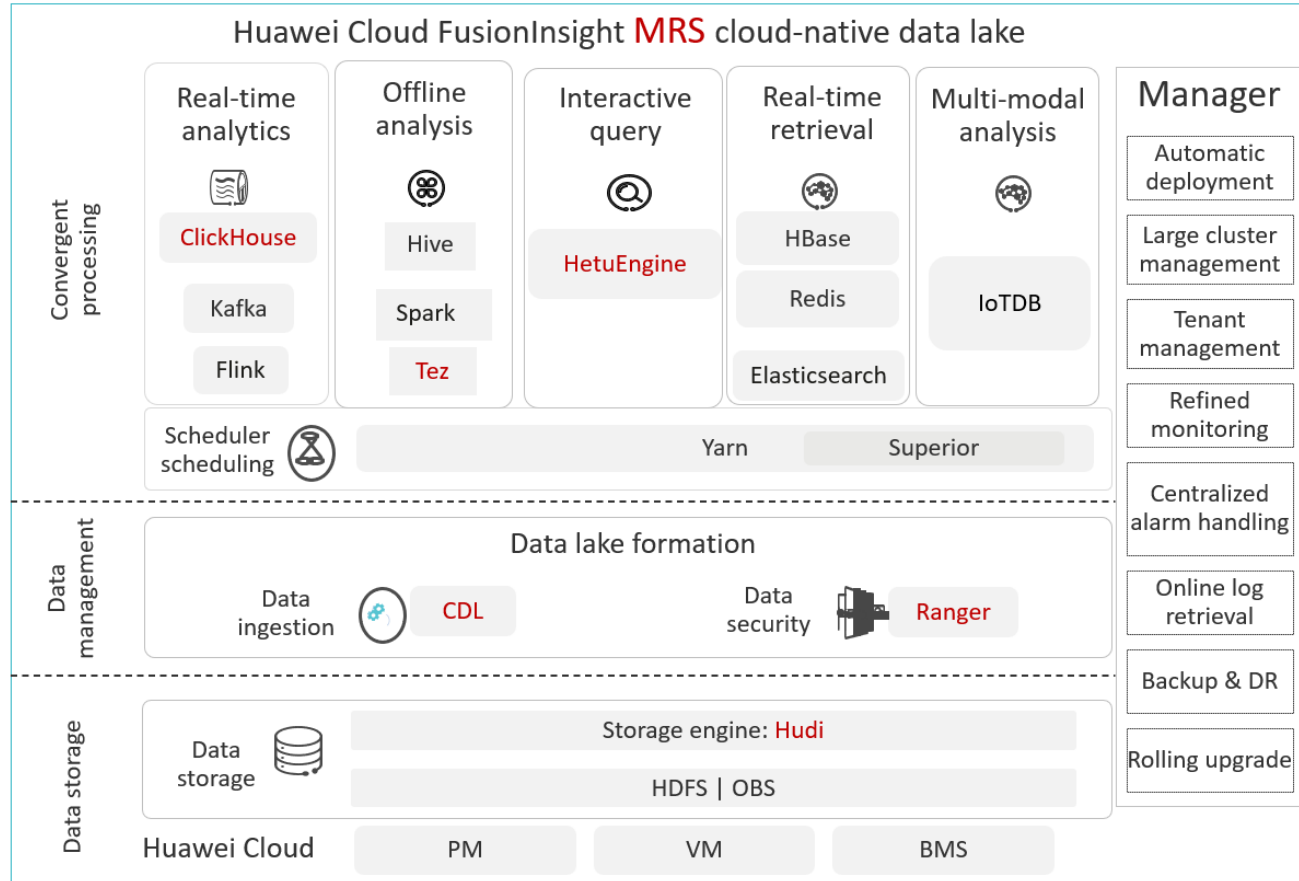
Cutting-edge open-source technologies

- In-depth reconstruction of mainstream engines, such as Spark, Hive, and Flink, with key technologies such as indexing, caching, and metadata; Huawei-developed CarbonData with millisecond-level point query; Superior Scheduler for 20,000+ nodes in a single cluster



High security and availability

- Cross-AZ HA of a single cluster, eliminating single points of failure (SPOFs), rolling patch installation/upgrade, task reconnection upon disconnection, and zero service interruption; multi-level security assurance capabilities, such as network resource isolation, account security, and data security control



Cloud-native architecture for agile building of data lakes

- Easy deployment: one-click cluster creation and service provisioning within 30 minutes
- Agile building: unified data import, metadata management, and security management for rapid construction of big data platforms
- Decoupled storage and compute: independent compute and storage resource expansion for convenient deployment



Three cloud-native data lakes with one architecture

- Offline data lake, real-time data lake, and logical data lake
- Specialized data marts: Data warehouse within a data lake, shortening the analysis link and construction period



Continuous evolution of enterprise-class versions

- Multi-engine converged analysis, improving analysis efficiency by 30%
- Up to 21,000 nodes are deployed in a cluster, supporting cluster federation
- Rolling upgrade for continuous evolution without service interruptions
- High availability

Chapter

10

MRS Application Scenarios



Offline data lake

- High-performance interactive query engine enables data processing within the lake.
- Unified metadata, global data visualization
- Converged analysis and unified SQL query are supported.



Real-time data lake

- High timeliness: Real-time incremental data import to the lake within seconds, from T+1 to T+0
- High resource utilization: Incremental data is scattered into the lake, meaning 2x more enhanced resource utilization
- Stream-batch convergence with unified SQL interfaces



Logical data lake

- Cross-lake, cross-domain, cross-warehouse, and all-domain data collaborative analysis
- Reduced data migration for computing without moving data, 50x higher analysis efficiency
- 10x faster service rollout efficiency (weeks -> days)



Abundant specialized data marts

- Lakes and marts in the same cluster for unified management and seamless interconnection
- Full self-service, millisecond-level real-time OLAP analysis of ClickHouse
- High-throughput, low-latency time series database IoTDB

The "**three data lakes + one data mart**" meets customers' requirements during different phases of data lake construction.

Overview of MRS

Components

- Hudi
- HetuEngine
- Ranger
- LDAP+Kerberos Security Authentication

Chapter 10

Content



Hudi is an open-source project launched by Apache in 2019 and became a top Apache project in 2020.

Huawei participated in Hudi community development in 2020 and used Hudi in FusionInsight.

Hudi is in a data lake table format, which provides the ability to update and delete data as well as consume new data on HDFS. It supports multiple compute engines and provides insert, update, and delete (IUD) interfaces and streaming primitives, including upsert and incremental pull, over datasets on HDFS.

Hudi is the file organization layer of the data lake. It manages Parquet files, provides data lake capabilities and IUD APIs, and supports compute engines.

Hudi Features

Supports fast updates.

Supports snapshot isolation for data write and query.

Manages file size and layout based on statistics.

Supports timeline.

Hudi Features

Supports data rollback.

Supports savepoints for data restoration.

Merges data asynchronously.

Optimizes data lake storage using the clustering mechanism

Overview of MRS

Components

- Hudi
- ▣HetuEngine
- ▣Ranger
- ▣LDAP+Kerberos Security Authentication

Chapter 10

Content



HetuEngine

HetuEngine is a self-developed high-performance engine for interactive SQL analysis and data virtualization. It seamlessly integrates with the big data ecosystem to implement interactive query of massive amounts of data in seconds. HetuEngine supports interactive quick query across sources (multiple data sources, such as Hive, HBase, Elasticsearch, and ClickHouse) and across domains (multiple regions or data centers).

Overview of MRS Components

- Hudi
- ▣HetuEngine
- ▣Ranger
- ▣LDAP+Kerberos Security Authentication

Chapter 10

Content



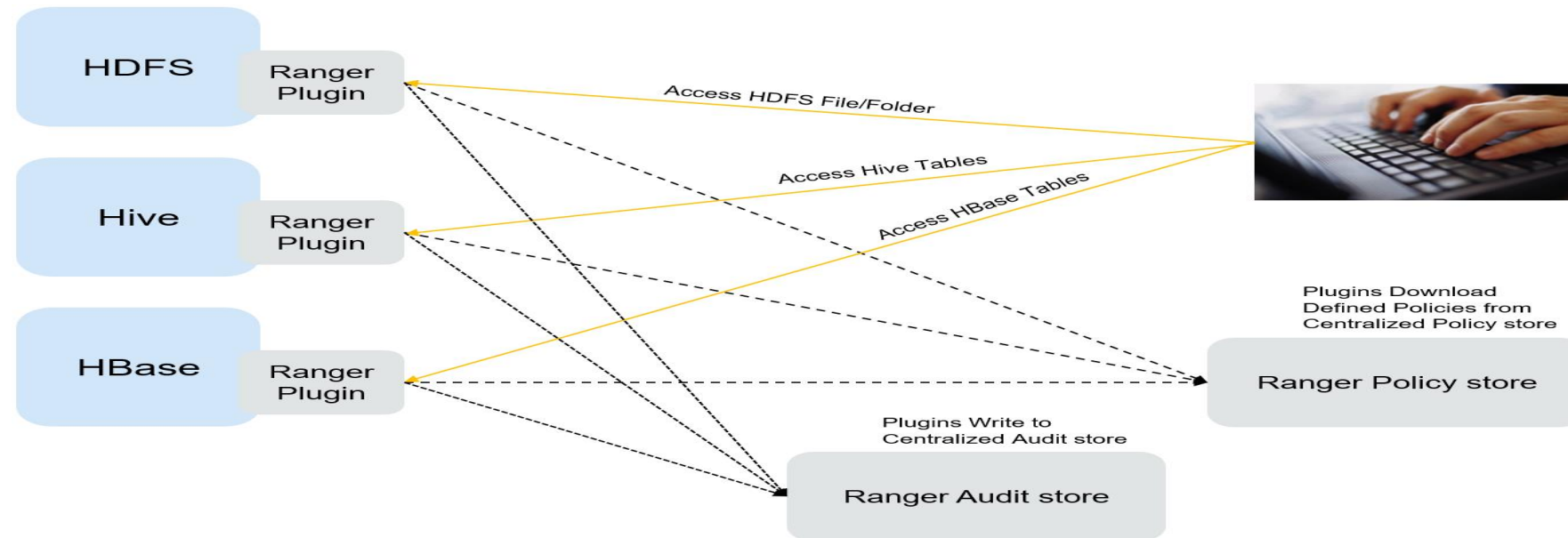
Ranger

Apache Ranger offers a centralized security management framework and supports unified authorization and auditing. It manages access control over Hadoop and related components, such as HDFS, Hive, HBase, Kafka, and Storm. Users can use the front-end web UI provided by Ranger to configure policies to control users' access to these components.

- Unified auditing
 - Ranger plug-ins can record audit logs. Currently, audit logs can be stored in local file systems.
- Reliability
 - Ranger supports two RangerAdmins working in active/active mode. Two RangerAdmins provide services at the same time. If either RangerAdmin is faulty, Ranger continues to work.
- High-performance
 - Ranger provides the load balancing capability. When a user accesses Ranger WebUI using a browser, the Load-Balance automatically selects the RangerAdmin with the lightest load to provide services.

Relationship Between Ranger and Other Components

Ranger provides PBAC authentication plug-ins for component servers. Currently, components like HDFS, Yarn, Hive, HBase, Kafka, Storm, and Spark2x support Ranger authentication. More components will become available in the future.



Overview of MRS

Components

- Hudi
- ▣HetuEngine
- ▣Ranger
- ▣LDAP+Kerberos Security Authentication

Chapter 10

Content



Lightweight directory access protocol (LDAP) is a protocol that helps users find data about organizations, persons, and more. LDAP has two main goals: to store data in the LDAP directory and authenticate users to access the directory. It also provides the communication language that applications require to send and receive information from directory services. A directory service provides access to *where* information on organizations, individuals, and other data is located within a network.

What Is LDAP Authentication?

- Here's a step-by-step breakdown of the authentication process:
- The client (an LDAP-ready system or application) sends a request to access information stored within an LDAP database.
- The client provides their LDAP server user credentials (username and password).
- The LDAP server cross-checks the user's submitted credentials against the core user identity data stored in its LDAP database.
- If the provided credentials match the stored core user identity, the client can access the requested information.
- Incorrect credentials will lead to denied access to the LDAP database.

Kerberos is a computer-network authentication protocol that works on the basis of tickets to allow nodes communicating over a non-secure network to prove their identity to one another in a secure manner. provides a centralized authentication server whose function is to authenticate users to servers and servers to users. In Kerberos Authentication server and database is used for client authentication. Kerberos runs as a third-party trusted server known as the Key Distribution Center (KDC). Each user and service on the network is a principal.

- **User Authentication:** User Authentication is one of the main applications of Kerberos. Users only have to input their username and password once with Kerberos to gain access to the network. The Kerberos server subsequently receives the encrypted authentication data and issues a ticket granting ticket (TGT).
- **Single Sign-On (SSO):** Kerberos offers a Single Sign-On (SSO) solution that enables users to log in once to access a variety of network resources. A user can access any network resource they have been authorized to use after being authenticated by the Kerberos server without having to provide their credentials again.
- **Authorization:** Kerberos also offers a system for authorization in addition to authentication. After being authenticated, a user can submit service tickets for certain network resources. Users can access just the resources they have been given permission to use thanks to information about their privileges

Huawei big data platform uses KrbServers to provide Kerberos functions for all components. To manage access control permissions on data and resources in a cluster, it is recommended that the cluster be installed in security mode. In security mode, a client application must be authenticated and a secure session must be established before the application can access resources in the cluster. MRS uses KrbServers to provide Kerberos authentication for all components, implementing a reliable authentication mechanism.

Service Authentication in the Cluster

In an MRS cluster in security mode, mutual access between services is implemented based on the Kerberos security architecture. When a service (such as HDFS) in the cluster is set to start, the corresponding sessionkey is obtained from Kerberos. If another service (such as Yarn) needs to access HDFS to add, delete, modify, or query data in HDFS, the corresponding TGT and ST must be obtained for secure access.

Application Development Authentication

MRS components provide application development interfaces for customers and upper-layer service product clusters. During application development, a cluster in security mode provides specified application development authentication interfaces to implement application security authentication and access.

Cross-Manager Mutual Trust

MRS provides the mutual trust function between two Managers to implement data read and write operations between them.

Overview of MRS
MRS Components
Huawei DataArts Studio
Data Governance
Huawei DataArts Studio

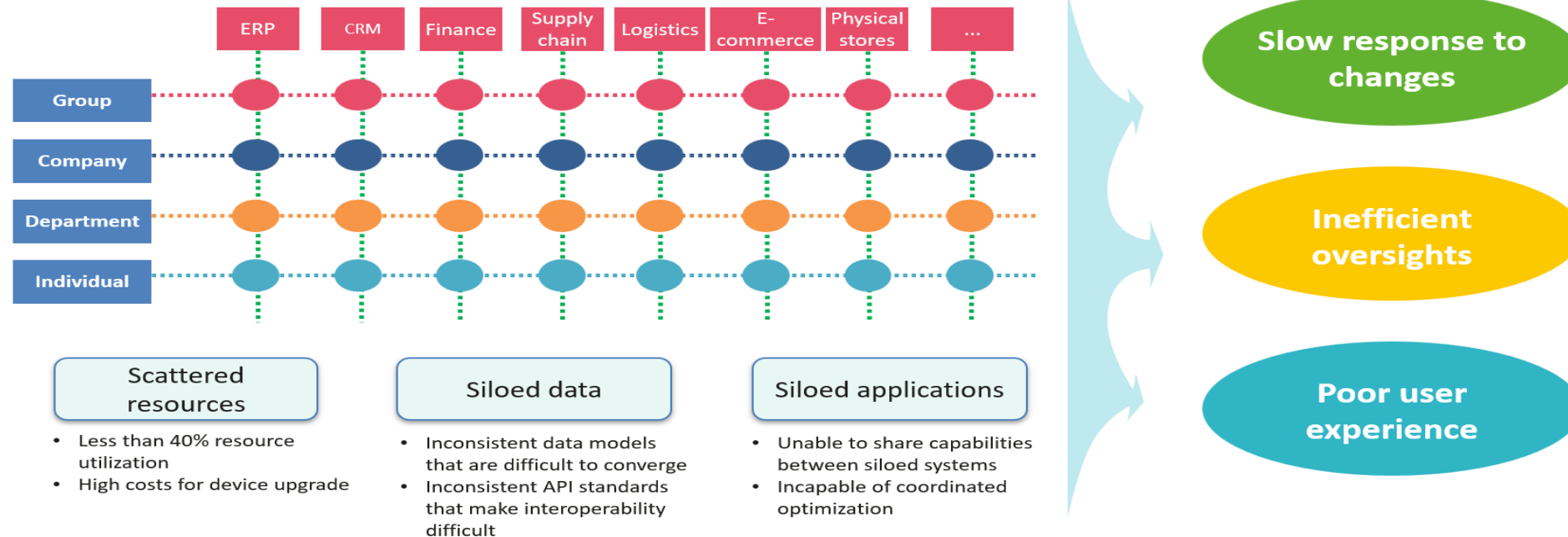
Chapter

10

Content



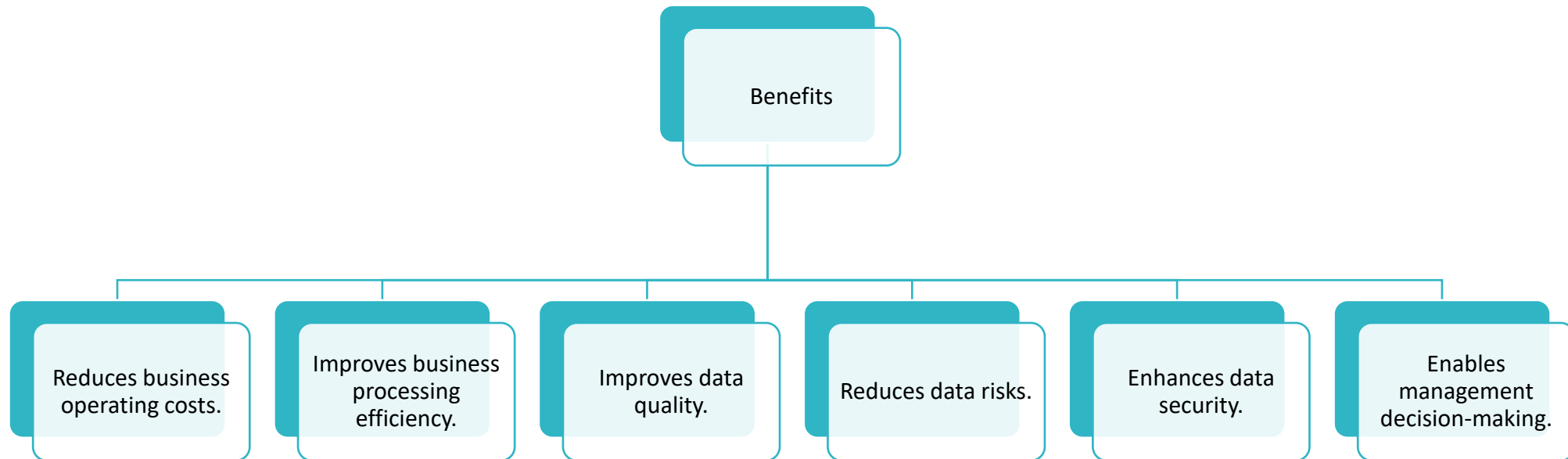
Application and Information Silos Severely Hinder Enterprises' Digital Development



The discipline of data governance treats data as enterprise assets and utilizes, protects, and optimizes those assets to maximize its value. This is done through the orchestration of people, processes, technologies, and policies within an enterprise.

Data governance aims to gain more value from data. It lays the foundation for enterprises to achieve their digital strategies. It is a management system consisting of organizations, regulations, processes, and tools.

Benefits of Data Governance



Overview of MRS
MRS Components
Huawei DataArts Studio
Data Governance
Huawei DataArts Studio

Chapter

10

Content



Huawei DataArts Studio

- DataArts Studio is a one-stop data governance and operations platform that provides intelligent data lifecycle management to help enterprises achieve digital transformation.
- It offers numerous functions, such as integrating and developing data, designing data standards, controlling data quality, managing data assets, creating data services, and ensuring data security. It supports the intelligent construction of industrial knowledge libraries and incorporates data foundations such as big data storage, computing, and analytical engines, helping enterprises easily build end-to-end intelligent data systems. These systems eliminate data silos, unify data standards, accelerate data monetization, and promote digital transformation.

DataArts Architecture incorporates data governance methods. You can use it to visualize data governance operations, connect data from different layers, formulate data standards, and generate data assets. You can standardize your data through ER modeling and dimensional modeling. Data Design is a good option for unified construction of metric platforms. With Data Design, you can build standard metric systems to eliminate data ambiguity and facilitate communications between different departments. In addition to unifying computing logic, you can also use it to query data and explore data value by subject.

- DataArts Quality can monitor your metrics and data quality, and screen out unqualified data in a timely manner.
- Business metric monitoring
 - You can use DataArts Quality to monitor the quality of data in your databases. You can create metrics, rules, or scenarios that meet your requirements and schedule them in real time or recursively.

- Data quality monitoring
 - You can create data quality rules to check whether the data in your databases is accurate in real time.
 - Qualified data must meet the following requirements: integrity, validity, timeliness, consistency, accuracy, and uniqueness. You can standardize data and periodically monitor data across columns, rows, and tables based on quality rules.

Chapter 10

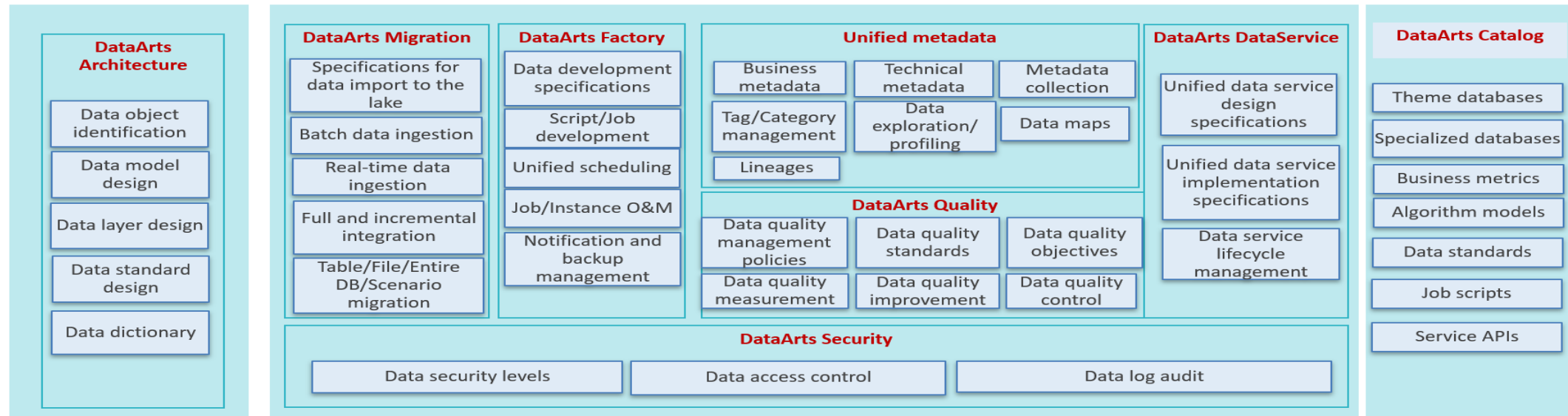
The Advantages of DataArts Studio

**Field-level lineage
analysis**

**Job management
DAG visualization**

**Data preview
50% higher efficiency**

**Privacy protection
Data masking | Audit**



Summary:

This chapter first described Huawei's big data platform, MRS, along with its advantages and application scenarios. It then went through some MRS components, including Hudi, HetuEngine, Ranger, and LDAP+Kerberos security authentication. Finally, this chapter introduced data governance and DataArts Studio (Huawei's data governance service).



مجموعة جامعة
الجميع الذكية
للتعليم القابضة

MIDOCEAN
UNIVERSITY
BY EVERYONE'S SMART UNIVERSITY



Thank You

