

Outline

- Understand data quality services in Microsoft SQL server.
- Install Data quality services (DSQ).
- Perform a data quality exercise on sample data.

Review from last class:

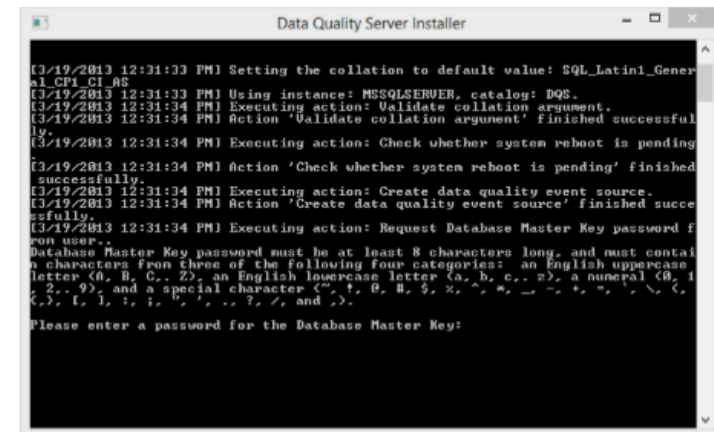
Traditional vs In memory computing

in memory -> needs lots of CPU to fold and unfold data

traditional needs disk space in order to store all the data

Installing DQS

- Requires Business Intelligence or Enterprise/Developer version of SQL Server 2012
- During SQL Server setup;
 - Instance Features -> Data Quality Services
 - Shared Features -> Data Quality Client
- Execute the Data Quality Server Installer;
 - C:\Program Files\Microsoft SQL Server\MSSQL11.MSSQLSERVER\MSSQL\Binn\DQSInstaller.exe
 - Data Quality Service – Data Quality Server Installer
(Apps - Microsoft SQL Server 2012)



```
Data Quality Server Installer

[3/19/2013 12:31:33 PM] Setting the collation to default value: SQL_Latin1_General_CP1_CI_AS
[3/19/2013 12:31:33 PM] Using instance: MSSQLSERVER, catalog: DQS.
[3/19/2013 12:31:34 PM] Executing action: Validate collation argument.
[3/19/2013 12:31:34 PM] Action 'Validate collation argument' finished successfully.
[3/19/2013 12:31:34 PM] Executing action: Check whether system reboot is pending successfully.
[3/19/2013 12:31:34 PM] Action 'Check whether system reboot is pending' finished successfully.
[3/19/2013 12:31:34 PM] Executing action: Create data quality event source.
[3/19/2013 12:31:34 PM] Action 'Create data quality event source' finished successfully.
[3/19/2013 12:31:34 PM] Executing action: Request Database Master Key password from user.
Database Master Key password must be at least 8 characters long, and must contain characters from three of the following four categories: an English uppercase letter (A, B, C, ..., Z), an English lowercase letter (a, b, c, ..., z), a numeral (0, 1, 2, ..., 9), and a special character (<\">, !, @, #, $, %, ^, *, _ , - , + , = , ' , ~ , ` , { , } , [ , ] , ; , : , \" , ' , , , / , & , and ,).
Please enter a password for the Database Master Key:
```

DQS Architecture

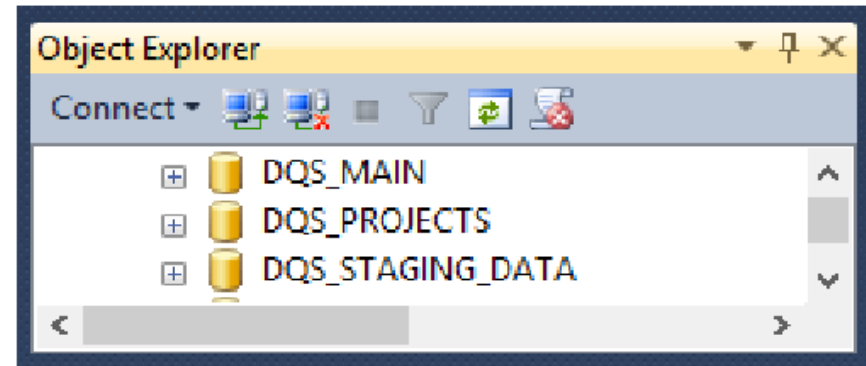
DQS Server

DQS Catalog (3 databases)

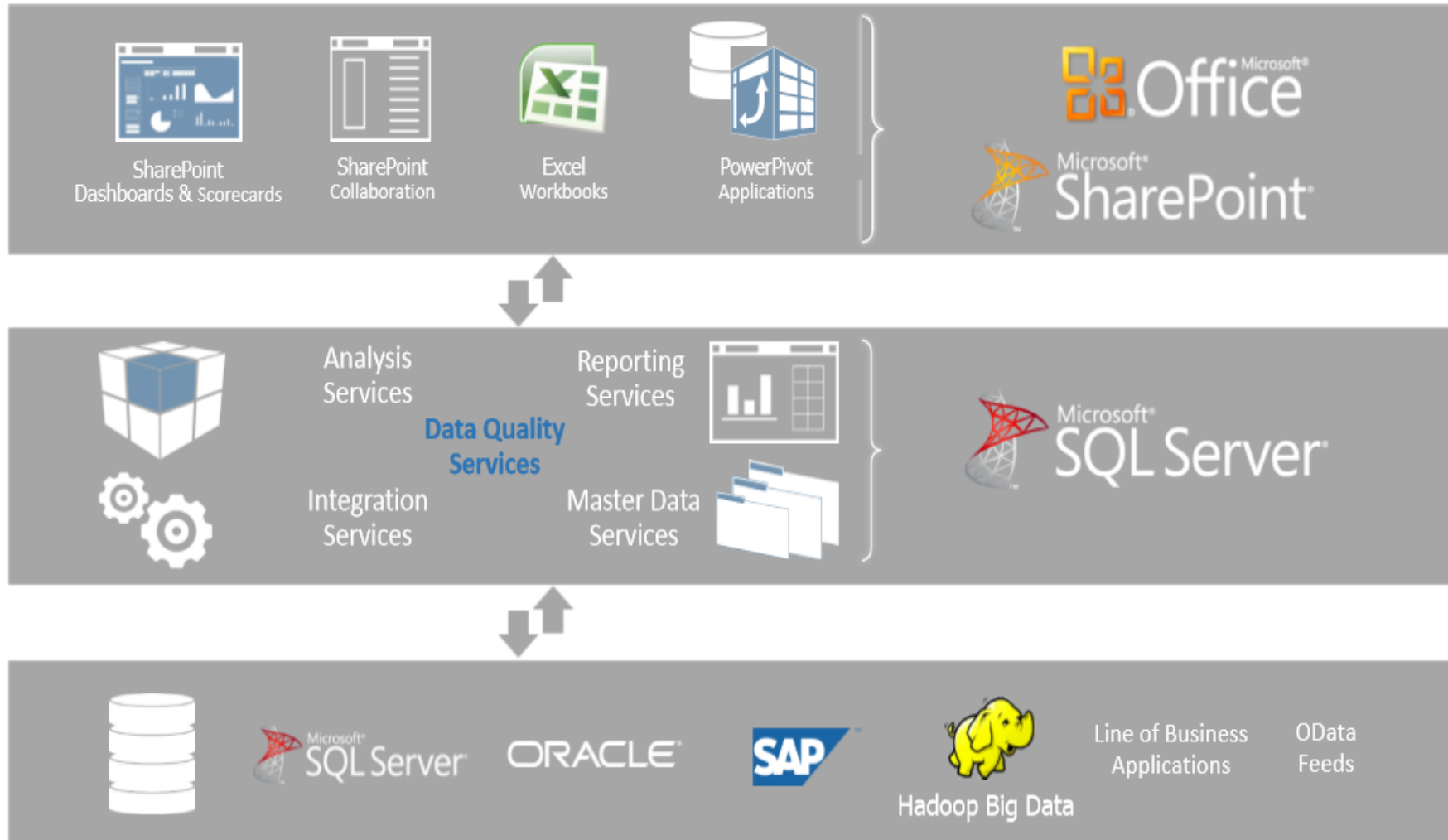
- DQS_MAIN (Knowledge Bases)
- DQS_PROJECTS (Projects)
- DQS_STAGING_DATA (Sandbox, scratch pad area)

Security – Database Roles

- dqs_administrator
- dqs_kb_editor
- dqs_kb_operator



Microsoft BI components



Dirty Data...

These are exam questions

- Do you have dirty data? (all projects have it! Its inevitable)

buy components services, HR buys one, supply chain buys another one. diff sets of data, etc.

Dirty Data

- Causes? Bad data entry Poor Data Governance Duplicate entities in different LOB (Line of Business) systems

Data Quality Issues

- Data quality issues can be divided into the following categories:
- Uniqueness - multiple copies of the same data, such as Bill Gates and Bill Geates
- Validity - Data is within established constraints
- Accuracy – Data is correct
- Standardization – Consistency in data representation
- Completeness – There are no missing pieces in the data.

Do you need Master Data Management (MDM)?

- Do you have instances of invalid data impacting business processes?
- Do you wish your business users could manage data themselves such as Customer and Product?
- Do you have IT resources spending time on data fixes and/or managing hierarchy definitions for the business?
- Do you have the need for data consolidation and the subsequent dissemination of the consolidated data to other systems?
- Do you have an environment of heterogeneous systems which all could benefit from a single view of domain data such as Customer or Product?

MDM Scenarios

Operational Data Management

Central data records management and consumption sourced by other operational systems

A company has adopted 6 new systems from a merger. The company needs the ability to propagate the correct customer information to each system in a consistent fashion. **MDS provides a platform for central schema, integration points and validation for Internal IT to develop a custom solution**

Data Warehouse Management (Analytical)

Enable business users to manage the dimensions and hierarchies of DW / Data Marts

Example: Business users utilize a data warehouse for reporting, but complain about the accuracy of the dimensions and lack of agility for updates. **MDS empowers the business users to manage dimensions themselves while IT can govern the changes**

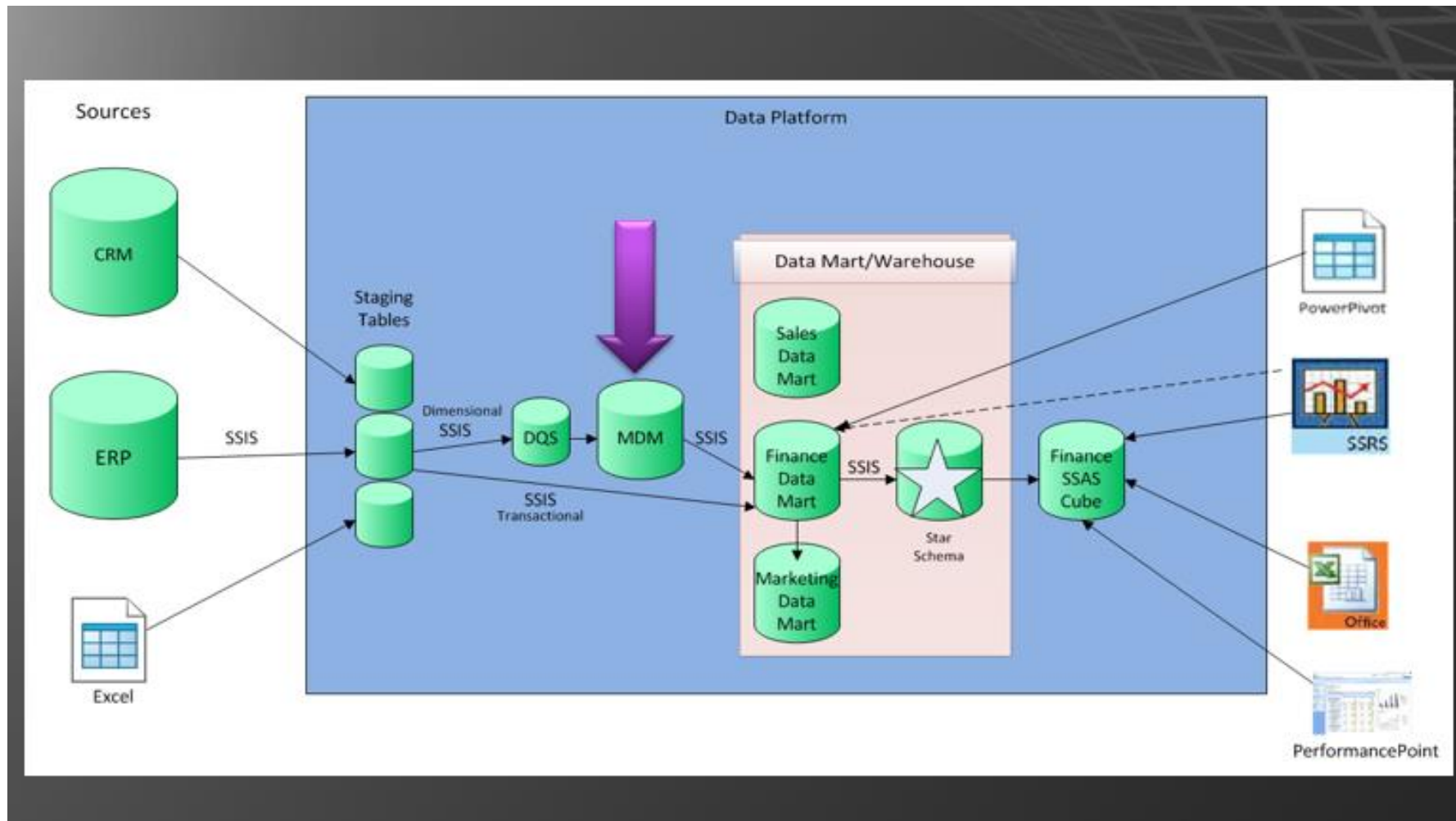
Data Solutions

Provides storage and management of the objects and metadata used as the application knowledge

- **Object mappings**
- **Reference Data**
- **Metadata management**

Example: Table A houses mapping data between two systems, and is also utilized by ETL processes for data transformation decisions. **MDS enables business users to manage the object mapping**

Where DQS / MDM fits in the Data Warehouse



Mis – Representation of Data

- Prospect in CRM System: Mark Smith | 613.111-1234 | Ottawa | ON | K1P 1K1
- Prospect buys goods now entered in POS System: Markus Smith | 1234 Stilton Ave | Kanata | ON | K1P 1K1
- Record also entered into Accounting System: Markus Smith | 1234 Stilton Avenue | Kanata | ON | K1P 1K1
- ETL process imports these records into the Data Warehouse / Data Mart
- FirstName LastName Phone Address City Province PostalCode
- Mark Smith 613.111-1234 Ottawa ON K1P 1K1
- Markus Smith 1234 Stilton Ave Kanata ON K1P 1K1
- Markus Smith 1234 Stilton Avenue Kanata ON K1P 1K1

Sample Data Representation

- Duplicate records and inaccurate, incomplete data

FirstName	LastName	Phone	Address	City	Province	PostalCode
Mark	Smith	613.111-1234		Ottawa	ON	K1P 1K1
Markus	Smith		1234 Stilton Ave	Kanata	ON	K1P 1K1
Markus	Smith		1234 Stilton Avenue	Kanata	ON	K1P 1K1

- What we want is a golden record (one version of the truth)

FirstName	LastName	Phone	Address	City	Province	PostalCode
Markus	Smith	613-111-1234	1234 Stilton Ave	Kanata	ON	K1P 1K1

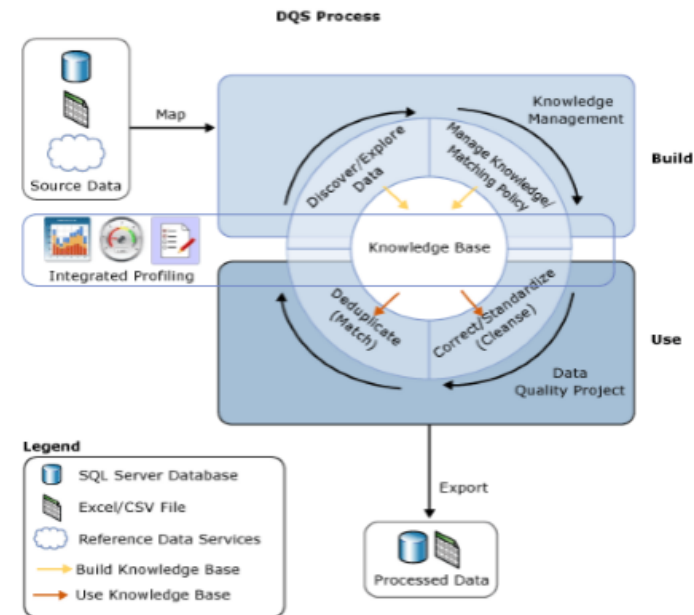
SQL Server Data Quality Services (DQS)

- Added to SQL Server 2012 enriching could mean adding in data to fill in missing fields
- Enables cleansing, matching, standardizing and enriching data
- Delivers trusted information for business intelligence, data warehouse, transaction processing workloads
- Knowledge-Driven Solution (create/edit)
- A knowledge management process that builds the knowledge base
- A data quality project that proposes changes to source data based on the knowledge in the knowledge base (cleansing and matching)
- A key component to an Enterprise Information Management (EIM) solution

Features of DQS

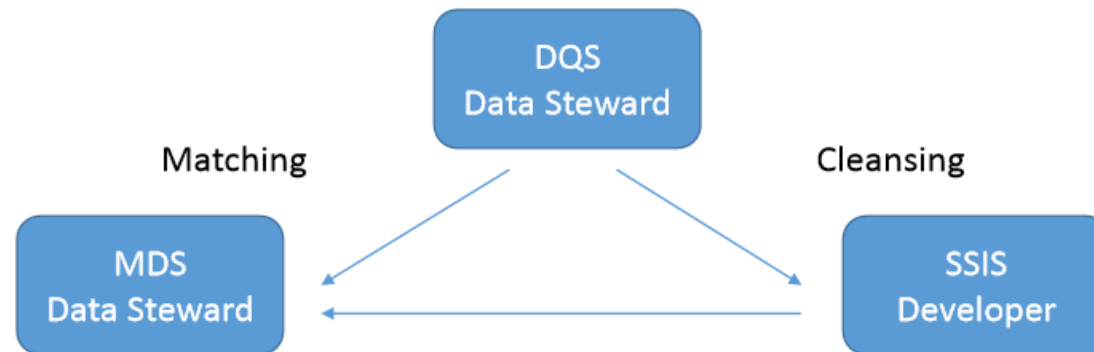
- DQS enables to resolve issues involving incompleteness, lack of conformity, inconsistency, inaccuracy, invalidity, and data duplication
- Provides the following features to resolve data quality issues:

- Data Cleansing
- Matching
- Reference Data Services
- Profiling
- Monitoring
- Knowledge Base



Data Steward

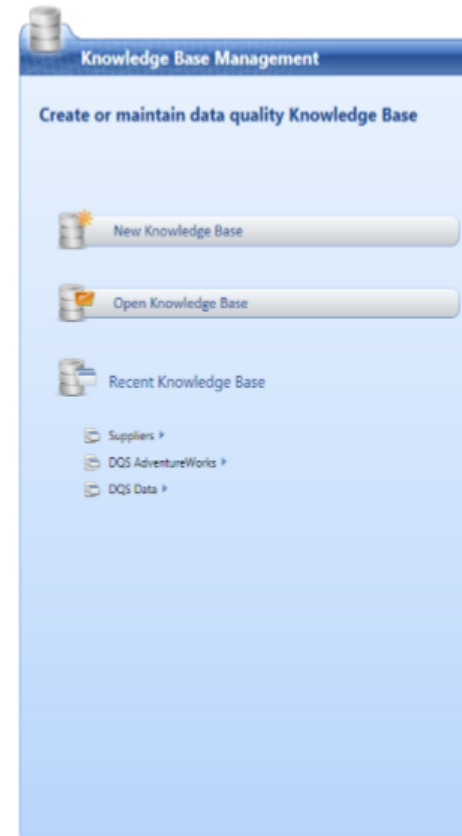
- **Key role** - Is usually a Business User and not from the Information Technology side
- Nutshell: Responsible for maintaining data elements in a metadata registry...
- Data Steward -> DQS Client
 - Create and edit Knowledge Bases
 - Run and process data though continually, iteratively, improving the Knowledge Bases
 - Knowledge Bases can be consumed and used by other Data Stewards and IT (SSIS / ETL Developers)



Knowledge Bases and Domains

The knowledge base is a repo of knowledge about your data that enables you to understand your data and maintain its integrity.

- Processes:
 - Computer-assisted
 - Interactive
- Components:
 - Knowledge Discovery
 - Domain Management
 - Reference Data Services
 - Matching Policy



Data Quality Projects

- Improve quality of source data by performing data cleansing and data matching activities using defined knowledge bases
- Cleansing Activity (2 step process)
- Computer-assisted : data is categorized (suggested, new, invalid, corrected, and correct)
- Interactive: data steward to approve, reject, or modify the proposed results from the computer-assisted cleansing process
- Matching Activity
- Using existing knowledge base matching policy
- Prevent and remove data duplication
- Data Profiling and Notifications
- Profiling provides data quality stats and info: completeness and accuracy
- Notification on actions that can be taken to enhance operations

Enterprise Information Management (EIM)

- SQL Server Data Quality Services (DQS) - Capture and record knowledge, rules, and actions
- SQL Server Master Data Services (MDS) - Master Data Management repository, Dimension data
- SQL Server Integration Services (SSIS) – Moves data, integration

