



Leveraging SAS for Efficient Data Warehousing

Dr. Michael Salé
Associate Professor of Management Information Systems
Meehan School of Business
Stonehill College

Boston Area SAS Users Group
Academic Ambassador

Agenda

- Introduction to Data Warehouses
- Importance of Data Warehouses in Business
- Role of SAS in Building Data Warehouses
- Understanding the ETL Process
- Data Sources for Your Data Warehouse
- Creating a Data Warehouse with a Star Schema
- Creating the Data Warehouse in SAS
- Scheduling SAS Data Warehouse Refreshes
- Enhancing BI Capabilities with Microsoft Power BI
- Review and Q&A Session



What is a Data Warehouse?

Definition:

- A data warehouse is a centralized repository designed to store integrated data from multiple sources. It supports reporting, analysis, and business intelligence operations by consolidating large volumes of data into a single, coherent database.

Key Characteristics:

- **Subject-Oriented:** Data in the warehouse is organized around major subjects, such as customers, products, or sales.
- **Integrated:** The data warehouse integrates data from disparate sources into a coherent dataset, ensuring consistency in naming conventions, measurement units, and data formats.
- **Non-Volatile:** Once data is entered into the warehouse, it is not changed or deleted.
- **Time-Variant:** The data warehouse contains historical data, enabling analyses based on time-series and trends.



A Data
Warehouse is...

The single version
of the truth!



Data Warehouses versus Databases



Purpose: Traditional databases are designed for day-to-day operations, focusing on transactional processing (OLTP - Online Transaction Processing), whereas data warehouses are designed for analysis and query processing (OLAP - Online Analytical Processing).



Data Organization: Databases often use a normalized structure to avoid data redundancy and ensure data integrity, which can slow down query processing. In contrast, data warehouses use denormalized structures, such as star and snowflake schemas, to speed up query performance.



Data Updates: In databases, data is frequently updated or deleted to reflect current operations. Data warehouses, on the other hand, accumulate historical data, growing over time without frequent updates to existing entries.

Importance of Data Warehouses to Business



Enhanced Decision Making and Business Intelligence



Data Consistency



Time Efficiency

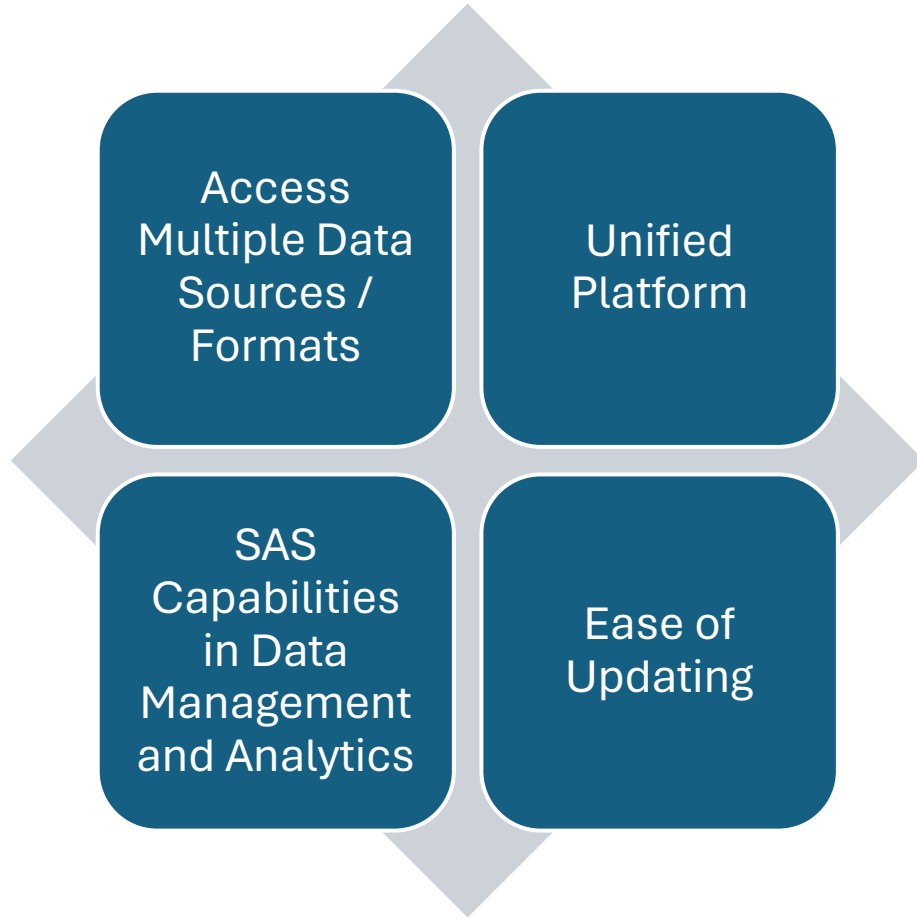


Historical Intelligence



Support for Large-Scale Analytics

What is SAS the Answer?



The ETL Process and Importance

ETL stands for **Extract, Transform, Load**, a process used in databases and data warehouses to make data useful for analysis. It's the backbone of data warehousing, enabling the efficient handling and preparation of data for insightful analysis and business intelligence.



Importance:

Data Quality and Consistency:

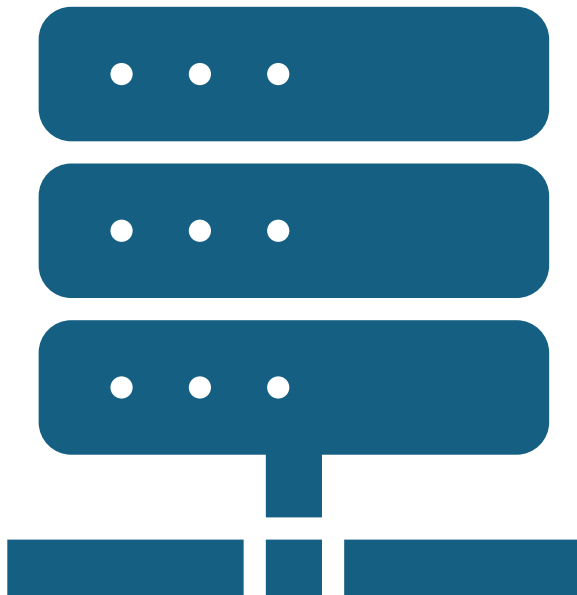
ETL processes are vital for maintaining high data quality and ensuring consistency across different data sources.

Performance Optimization:

By preprocessing data, ETL operations can significantly enhance the performance of data retrieval and analysis tasks.

Flexibility and Scalability:

Well-designed ETL processes can adapt to changes in data sources and structures, supporting business growth and evolution.



The ETL Process – Extract

The first step involves collecting or extracting data from various internal and external sources, such as operational databases, flat files, external data streams, and cloud sources.

- **Key Points:**
 - Data is often raw and can include structured, semi-structured, or unstructured formats.
 - The extraction process must ensure data integrity and minimize impact on source systems, often using techniques like incremental extraction.



The ETL Process – Transform

This phase involves cleaning, standardizing, deduplicating, and converting extracted data into a format suitable for analysis and reporting.

- **Key Points:**
 - Transformation rules might include aggregation, normalization, key generation, and value mapping.
 - It's crucial for improving data quality and ensuring compatibility with the target data warehouse schema.



The ETL Process – Load

The final step is loading the transformed data into the data warehouse or another destination system for querying and analysis.

- **Key Points:**
 - Loading can be done in batches (batch loading) or in real-time (streaming) using database triggers.
 - Consideration for the load process includes handling data volume without affecting the performance of the target system.

Data Sources for Your Data Warehouse



DATABASE TABLES
(MS SQL SERVER,
ORACLE, ETC.)



FLAT FILES
(.CSV, .TXT)



SAS DATA SETS
(.SAS7BDAT)

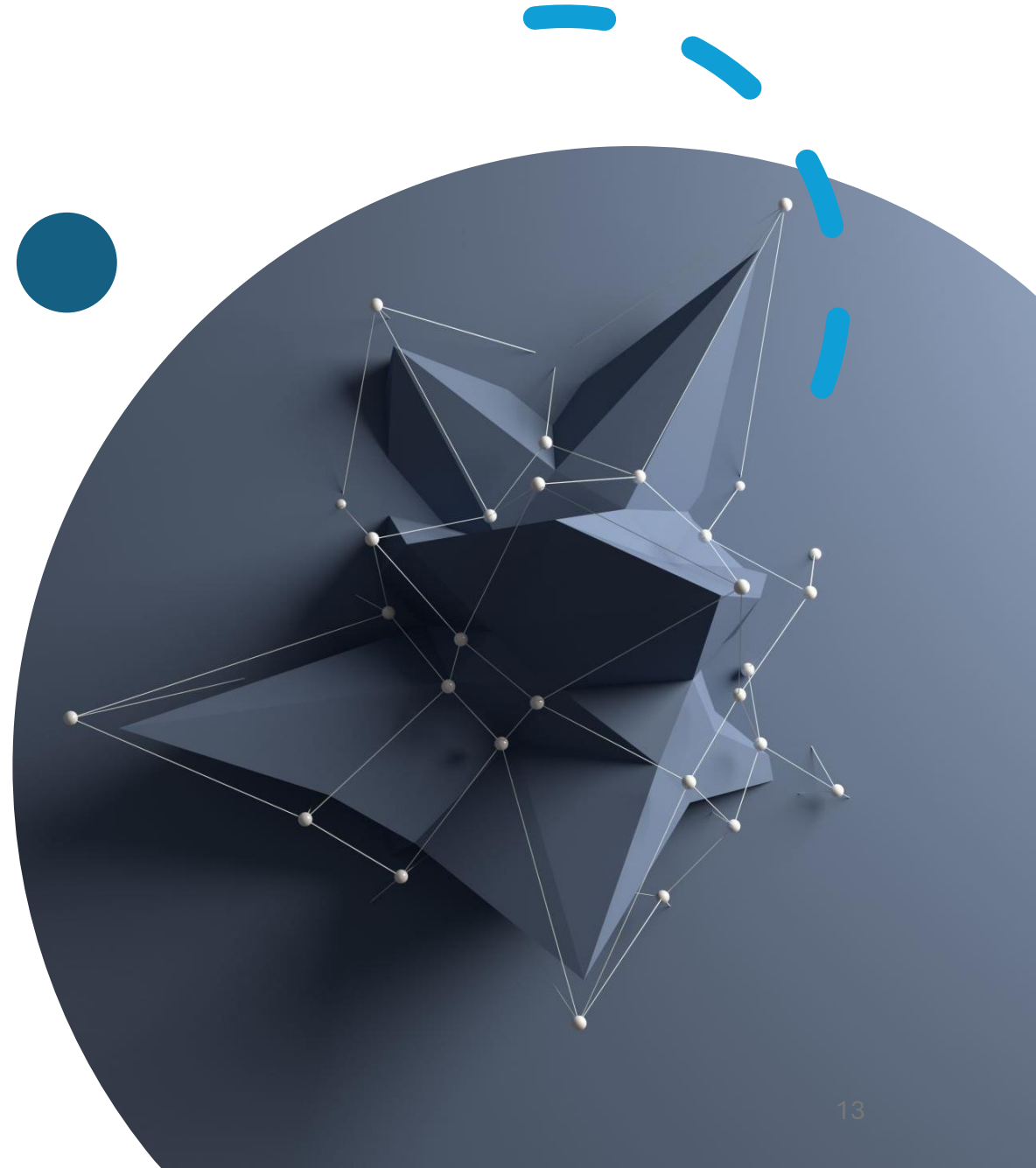


WEB SERVICE APIS
(JSON, XML)

Star Schema

The star schema is a fundamental database architecture for data warehouses that simplifies complex data structures into a format optimal for querying and report generation. It consists of a central fact table surrounded by dimension tables.

- **Components of a Star Schema:**
 - Fact Tables
 - Dimension Tables



Fact Tables

The fact table is the core of the star schema, containing quantitative data for analysis and reporting, such as sales revenue or units sold.

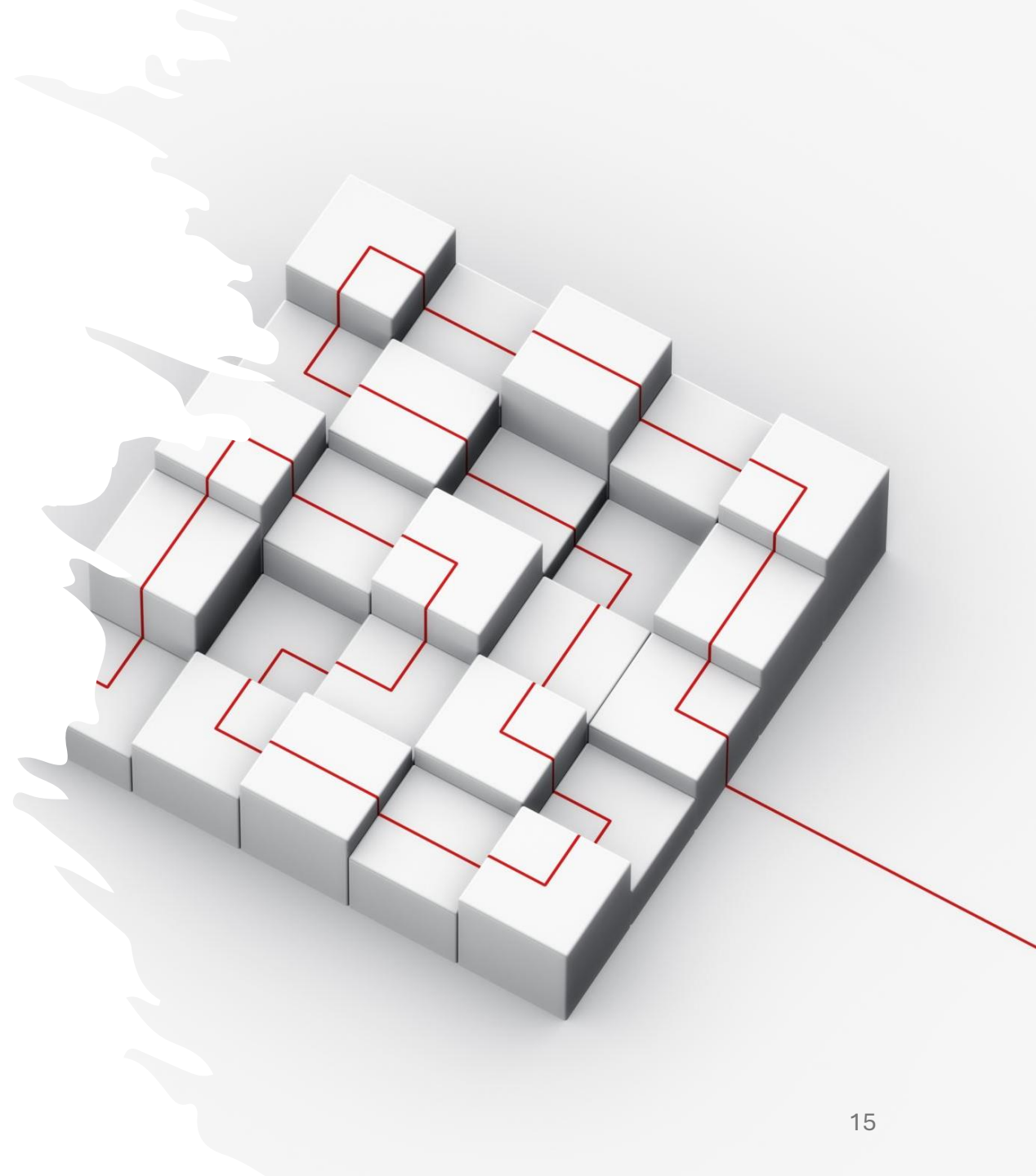
- **Key Features:**
 - Fact tables store metrics and foreign keys from dimension tables, facilitating the connection of different data dimensions.



Dimension Tables

Dimension tables store descriptive attributes related to the dimensions of the business process, such as time, product, customer, and location.

- **Key Features:**
 - These tables help in filtering, grouping, or labeling facts, making data more accessible and understandable.



Designing a Star Schema

1

Identify Business Process:

- Start by identifying the business process or analysis focus (e.g., sales performance).

2

Determine Facts:

- Define the measurable quantities (facts) that support the business process.

3

Define Dimensions:

- Identify the descriptive attributes (dimensions) relevant to the facts.

4

Design Tables:

- Create the fact table with foreign keys to dimension tables and populate dimension tables with descriptive attributes.

Schedule Data Warehouse Updates / Refreshes

You can update / refresh your data warehouse in two ways:

- **Batch Updates**
using Windows Task Scheduler
- **Real-Time (stream feed)**
using database triggers

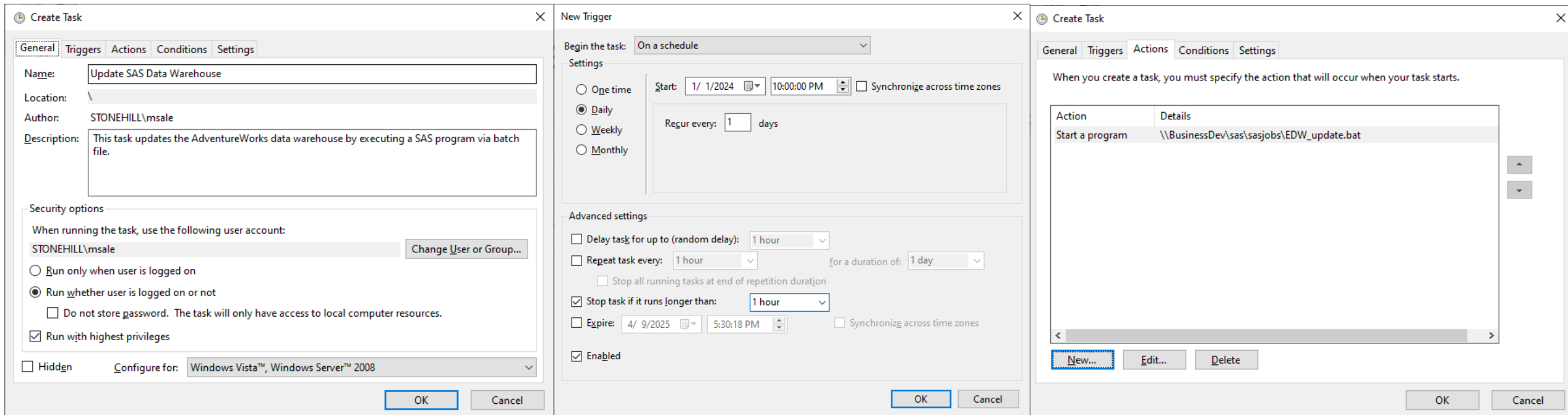


Windows Task Scheduler

Sample Batch File

```
"C:\Program Files\SASHome\SASFoundation\9.4\sas.exe" -SYSIN \\BusinessDev\sas\sasjobs\EDW.sas
```

Task Scheduler Screens



The image displays three screenshots of the Windows Task Scheduler interface, illustrating the configuration of a task.

Left Screenshot: Create Task (General tab)

- Name:** Update SAS Data Warehouse
- Location:** \
- Author:** STONEHILL\msale
- Description:** This task updates the AdventureWorks data warehouse by executing a SAS program via batch file.
- Security options:** When running the task, use the following user account: STONEHILL\msale. Run whether user is logged on or not. Run with highest privileges.
- Hidden:**
- Configure for:** Windows Vista™, Windows Server™ 2008

Middle Screenshot: New Trigger

- Begin the task:** On a schedule
- Settings:** Daily. Start: 1/ 1/2024 10:00:00 PM. Recur every: 1 days.
- Advanced settings:** Stop task if it runs longer than: 1 hour. Enabled.

Right Screenshot: Create Task (Action tab)

- Action:** Start a program
- Details:** \\BusinessDev\sas\sasjobs\EDW_update.bat

Database Trigger

1

Define the Trigger

- First, create a trigger on the table where rows might be updated.

2

Use SQL Server Agent

- Since direct execution is not feasible, use SQL Server Agent to periodically check the logging table for new entries and execute the batch file accordingly.

```
CREATE TRIGGER trgAfterUpdate
ON YourTableName
AFTER UPDATE
AS
BEGIN
    -- Set NeedsProcessing to 1 for the updated rows
    UPDATE YourTableName
    SET NeedsProcessing = 1
    FROM Inserted i
    WHERE YourTableName.ID = i.ID;
END;
GO
```

•**Security and Permissions:** Running external scripts or applications from SQL Server can introduce security risks. Ensure you tightly control and monitor who has access to create or modify the SQL jobs, triggers, and scripts.

•**Performance:** Frequent checks or complex scripts can impact database and server performance.



Review

- Defined data warehouses and their importance in business
- Discussed why SAS is the perfect platform for creating data warehouses
- Reviewed different data sources and the ETL process
- Designed a data warehouse using the star schema method
- Created our fact and dimension tables using SAS
- Enhanced our BI capabilities using Power BI
- Scheduled our data warehouse updates in batch and via triggers





Questions

- Michael Salé
Associate Professor of Management
Information Systems
Meehan School of Business
Stonehill College
msale@stonehill.edu