**Assignment 7 Celebal Technologies**

**Solution 1:**

**1. Processing "CUST\_MSTR" Files**

**File Naming Convention Example:** CUST\_MSTR\_20191112.csv

**Transformation Needed:**

* Extract date 20191112 from the filename.
* Convert it to the format 2019-11-12.
* Add this extracted date as an additional column Date in the output.
* Load data into the CUST\_MSTR table.

**Steps:**

* Read each CUST\_MSTR\_\* file from the data lake container.
* Extract the date part (20191112) from the filename.
* Convert this date to the format YYYY-MM-DD (2019-11-12).
* Append this transformed date as a new column Date to the dataset.
* Load the processed dataset into the CUST\_MSTR table in your database.

**2. Processing "master\_child\_export" Files**

**File Naming Convention Example:** master\_child\_export-20191112.csv

**Transformations Needed:**

* Extract date 20191112 from the filename.
* Convert it to the format 2019-11-12.
* Add this extracted date as an additional column Date in the output.
* Convert the date 20191112 to 20191112 (same format, but as a number for a key).
* Add this date key as an additional column DateKey.
* Load data into the master\_child table.

**Steps:**

* Read each master\_child\_export-\* file from the data lake container.
* Extract the date part (20191112) from the filename.
* Convert this date to the format YYYY-MM-DD (2019-11-12) and store it as Date.
* Convert the same date (20191112) to a number format (20191112) and store it as DateKey.
* Load the processed dataset into the master\_child table in your database.

**3. Loading "H\_ECOM\_ORDER" Files**

**File Naming Convention Example:** H\_ECOM\_ORDER.csv

**Direct Loading:**

* Load the file directly into the H\_ECOM\_Orders table in your database without any modifications.

**Steps:**

* Read H\_ECOM\_ORDER.csv from the data lake container.
* Load the dataset into the H\_ECOM\_Orders table in your database.

**Implementation Considerations:**

* **File Handling:** Iterate through files in the data lake container based on file patterns (CUST\_MSTR\_\*, master\_child\_export-\*, H\_ECOM\_ORDER.csv).
* **Data Extraction:** Extract date information from filenames using string manipulation or regular expressions.
* **Data Transformation:** Convert extracted date formats as specified (YYYY-MM-DD and YYYYMMDD).
* **Database Operations:** Ensure truncate-load operations for daily processing to replace existing data with new data.
* **Automation:** Implement a scheduled process (e.g., using cron jobs or Azure Data Factory pipelines) to automate this daily data ingestion and transformation.

**Solution2:**

Creating a bill of materials (BoM) for the Azure resources needed to handle the described data sources involves estimating the necessary compute, storage, and integration components. Here’s a high-level breakdown of what you might consider:

**Data Sources and Requirements:**

1. **Oracle (On-Premise)**:
   * **Monthly Incremental Data Size**: 30 GB
   * **Total Table Count**: 20
2. **Salesforce**:
   * **Monthly Incremental Data Size**: 50 GB
   * **Total Table Count**: 120
3. **Semi-structured files on FTP**:
   * **Monthly Data Size**: 5 GB
   * **Approximate File Count per Month**: 20

**Azure Bill of Materials (BoM):**

**1. Data Ingestion:**

* **Azure Data Factory (ADF)**:
  + Use ADF to orchestrate data movement and transformation workflows.
  + Pricing based on activities and data volume processed.
  + **Link**: [Azure Data Factory Pricing](https://azure.microsoft.com/en-us/pricing/details/data-factory/)

**2. Storage:**

* **Azure Blob Storage**:
  + Store raw and processed data.
  + Consider hot, cool, or archive tiers based on access frequency.
  + **Link**: [Azure Blob Storage Pricing](https://azure.microsoft.com/en-us/pricing/details/storage/blobs/)

**3. Compute:**

* **Azure Virtual Machines (VMs)**:
  + For running integration runtime for ADF or processing tasks.
  + Choose VM sizes based on workload requirements (CPU, memory).
  + **Link**: [Azure VM Pricing](https://azure.microsoft.com/en-us/pricing/details/virtual-machines/)

**4. Database:**

* **Azure SQL Database** (for structured data from Oracle and Salesforce):
  + Choose the appropriate service tier based on performance and storage requirements.
  + **Link**: [Azure SQL Database Pricing](https://azure.microsoft.com/en-us/pricing/details/sql-database/)
* **Azure Cosmos DB** (for semi-structured data from FTP):
  + NoSQL database suitable for JSON data.
  + Choose consistency levels based on application needs.
  + **Link**: [Azure Cosmos DB Pricing](https://azure.microsoft.com/en-us/pricing/details/cosmos-db/)

**5. Networking:**

* **Azure Virtual Network (VNet)**:
  + Securely connect Azure resources to on-premises and other cloud resources.
  + **Link**: [Azure VNet Pricing](https://azure.microsoft.com/en-us/pricing/details/virtual-network/)

**6. Monitoring and Management:**

* **Azure Monitor**:
  + Monitor the health and performance of Azure resources.
  + **Link**: [Azure Monitor Pricing](https://azure.microsoft.com/en-us/pricing/details/monitor/)

**Additional Considerations:**

* **Azure Cost Management**:
  + Monitor, allocate, and optimize costs across Azure.
  + **Link**: [Azure Cost Management](https://azure.microsoft.com/en-us/pricing/details/cost-management/)