Neo Stats – Data Engineer Internship Case Study

Customer 360 Analytics Platform

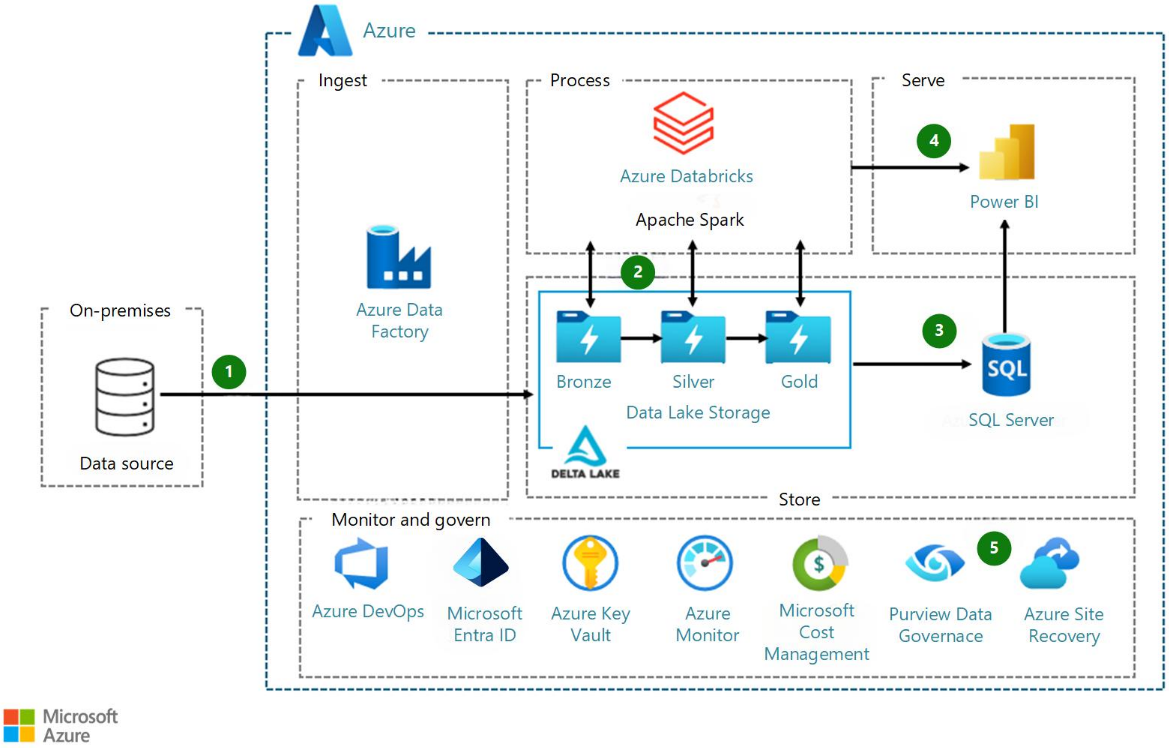
Submitted by: Melapakula Niranjan

Roll No: 2022BCSE07AED384

Role: Data Engineer Intern

Tools: Azure, Databricks, SQL, Power BI

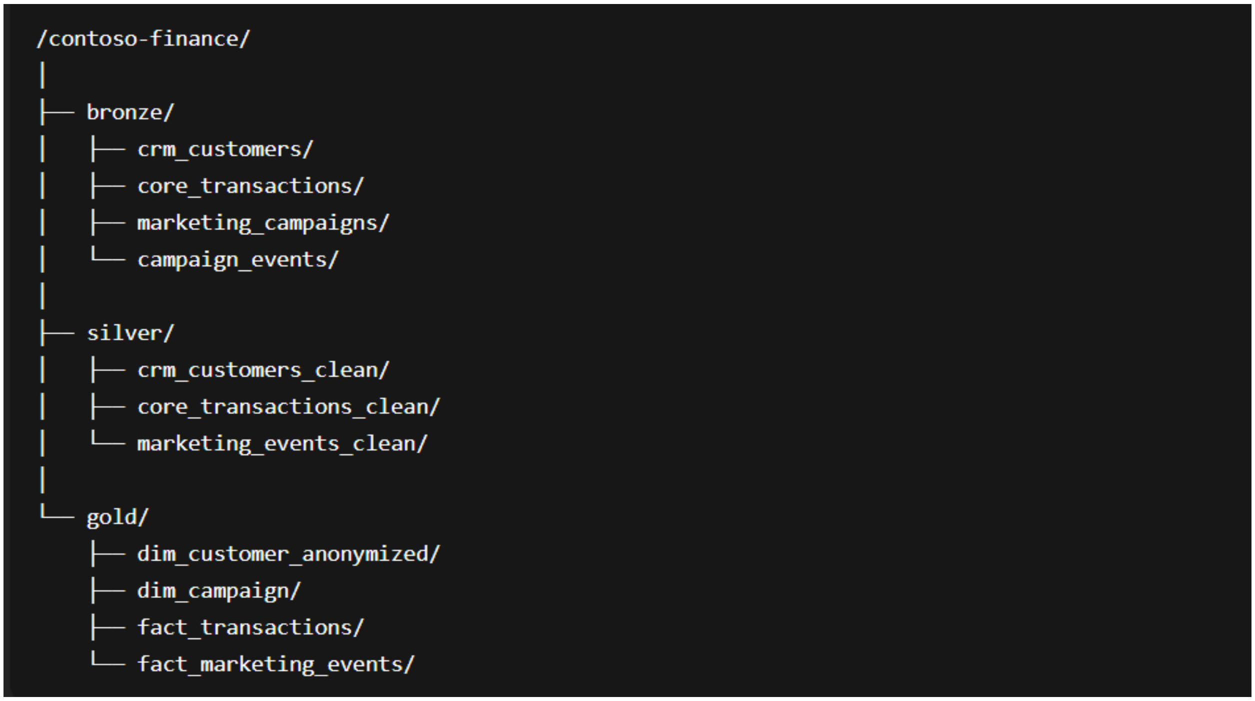
Contoso Finance – Customer 360 Analytics Platform on Azure :



Architecture Overview :

This solution implements a secure Azure lakehouse architecture to support Customer 360 analytics for Contoso Finance while enforcing strict PII governance. Data from CRM, transaction, and marketing sources is ingested using Azure Data Factory and stored in Azure Data Lake Gen2 following a Bronze – Silver – Gold layering approach. Azure Databricks processes raw into standardized and curated datasets, applying data quality checks and transforming sensitive customer attributes. Direct PII is retained only in restricted layers, while the Gold layer exposes fully anonymized and business-ready fact and dimensions tables designed for analytics. Azure key Vault manages secrets and encryption keys, and Microsoft Purview provides data classification and lineage. Power BI consumes data exclusively from the Gold Layer, with optional SQL endpoints used for optimized reporting, ensuring secure self-service analytics without exposing customer PII.

Azure Data Lake Folder Structure (Bronze / Silver / Gold):



```
/contoso-finance/
|
├── bronze/
|   ├── crm_customers/
|   ├── core_transactions/
|   ├── marketing_campaigns/
|   └── campaign_events/
|
├── silver/
|   ├── crm_customers_clean/
|   ├── core_transactions_clean/
|   └── marketing_events_clean/
|
└── gold/
    ├── dim_customer_anonymized/
    ├── dim_campaign/
    ├── fact_transactions/
    └── fact_marketing_events/
```

ADLS Gen 2 Organization Strategy:

Azure Data Lake Storage Gen2 is organized using a Bronze – Silver – Gold Structure to logically separate raw, processed, and curated data while enforcing data governance and access control.

The Bronze layer stores raw CRM, transaction, and marketing data exactly as ingested from sources systems. This layer preserves historical records or traceability and reprocessing and contains sensitive PII, therefore access is strictly limited to data engineering roles.

The silver layer holds cleaned and standardized datasets where schema validation, deduplication, and data quality checks are applied. Data is made analytics-consistent but still treated as sensitive.

The Gold Layer contains curated, business-ready fact and dimensions tables designed or customers 360 analytics. All direct PII is removed or transformed, and customers are represented using anonymized surrogate keys. This layer is the only layer exposed to Power BI and business users.

Gold Layer Tables Overview:

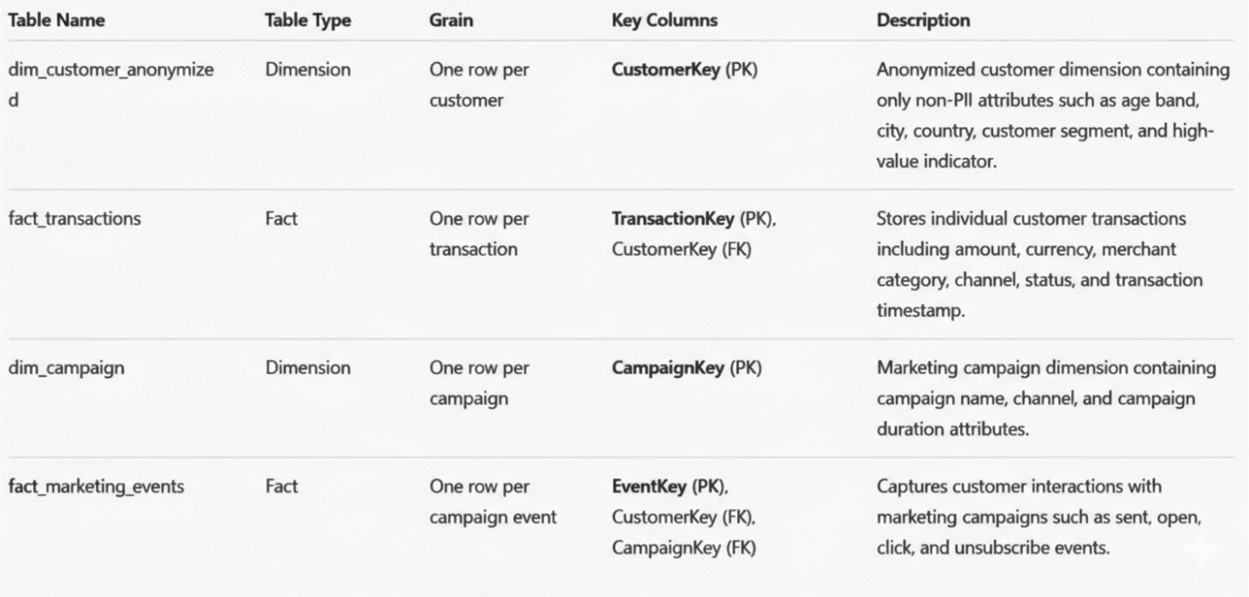
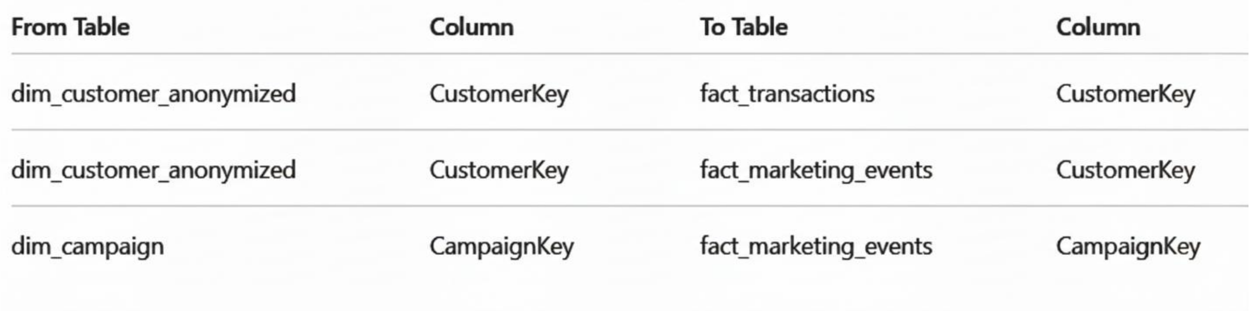


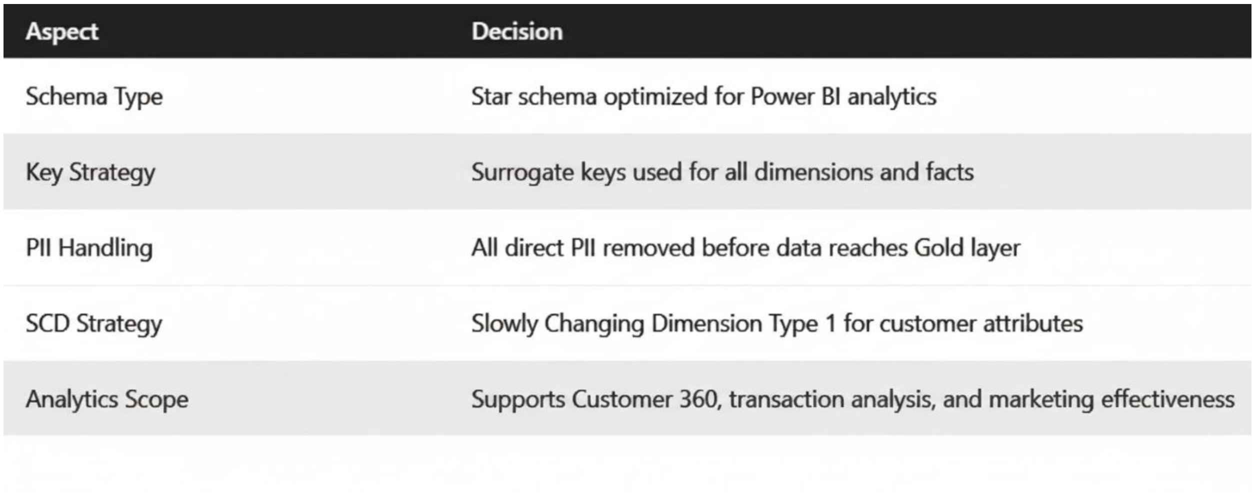
Table Name	Table Type	Grain	Key Columns	Description
dim_customer_anonymized	Dimension	One row per customer	**CustomerKey** (PK)	Anonymized customer dimension containing only non-PII attributes such as age band, city, country, customer segment, and high-value indicator.
fact_transactions	Fact	One row per transaction	**TransactionKey** (PK), CustomerKey (FK)	Stores individual customer transactions including amount, currency, merchant category, channel, status, and transaction timestamp.
dim_campaign	Dimension	One row per campaign	**CampaignKey** (PK)	Marketing campaign dimension containing campaign name, channel, and campaign duration attributes.
fact_marketing_events	Fact	One row per campaign event	**EventKey** (PK), CustomerKey (FK), CampaignKey (FK)	Captures customer interactions with marketing campaigns such as sent, open, click, and unsubscribe events.

Key Relationships:



From Table	Column	To Table	Column
dim_customer_anonymized	CustomerKey	fact_transactions	CustomerKey
dim_customer_anonymized	CustomerKey	fact_marketing_events	CustomerKey
dim_campaign	CampaignKey	fact_marketing_events	CampaignKey

Design & Privacy Considerations:



Aspect	Decision
Schema Type	Star schema optimized for Power BI analytics
Key Strategy	Surrogate keys used for all dimensions and facts
PII Handling	All direct PII removed before data reaches Gold layer
SCD Strategy	Slowly Changing Dimension Type 1 for customer attributes
Analytics Scope	Supports Customer 360, transaction analysis, and marketing effectiveness

PII Identification and Classification:

The CRM customer dataset contains multiple categories of personally identifiable information (PII). Direct identifiers include customer_id, first name, last name, email address, phone number, and address fields. Sensitive PII includes data of birth, while attributes such as city and country are treated as quasi-identifiers. Transactional and marketing datasets do not contain direct PII and are considered non sensitive.

Customer De-Identification Strategy:

To prevent exposure of sensitive customer data, all direct PII is retained only in restricted raw layers. In the Gold layer, customers are represented using a generated surrogate key (Customer Key) derived from the source customer_id through hashing or tokenization. Attributes such as date of birth are transformed into derived values like age or age bands, while names, contact details, and full addresses are completely removed. This ensures that analytics users cannot directly or indirectly identify individual customers

Role-Based Access Control Model:

Access to data is governed using role-based controls. Data Engineers and platform administrators have audited access to all layers, including raw PII, for operational and troubleshooting purposes. Data Scientists are granted access to processed datasets with masked identifiers and non-sensitive attributes. Business users and Power BI consumers are restricted to the gold layer only, which contains fully anonymized and analytics-ready data.

Governance and Security Implementation:

Microsoft Purview is used to classify sensitive columns, track data lineage from source to reporting, and provide a centralized data catalog. Azure Key Vault securely stores secrets, encryption keys, and hashing salts used during anonymization. Azure Data Lake Storage access controls and audit logs ensure that all sensitive data access is monitored and complaint with privacy requirements.

Incremental Data Ingestion Approach

Incremental ingestion is designed to efficiently process only new or changed data while maintaining data consistency. For the CRM customer dataset, the updated_at timestamp is used to identify newly created or modified records. During each pipeline run, only records with an updated_at value greater than last processed watermark are ingested and merged into the curated customer dimension.

Core transactions data is treated as append-only. New transaction files are ingested on a daily basis, and only previously unprocessed records are loaded into the Silver and Gold layers. This approach avoids reprocessing historical data improves pipeline performance.

Customer Dimension Upsert Strategy

The anonymized customer dimension follows a Slowly Changing Dimension (SCD) Type 1 approach. When a customer record changes, the existing record in the Gold layer is updated in place using the generated CustomerKey. This strategy simplifies data management and aligns with privacy requirements by avoiding the retention of historical sensitive attributes.

Transaction Fact Load Strategy

Transaction records are incrementally appended to the fact table based on Transaction timestamp or file ingestion dates. Duplicate transactions are prevented through primary key validation at ingestion time. Referential integrity is enforced to ensure that all transactions reference a valid anonymized customer.

Data Quality Controls

Data quality checks are applied during Silver layer processing to ensure reliability of analytics output. Key Validations include enforcing uniqueness of customer identifiers, ensuring transaction amounts are non-null and non-negative, validating that transaction timestamps are not in the future, confirming referential integrity between customers and transactions, and restricted layers. Records failing quality checks are logged and excluded from downstream consumption.

Note: This solution was independently designed for the purpose of this case study.