

SQL Data Warehouse, Analytics, & Reporting Project

Presented by: Nitin Kunigal

Agenda

- Business Context & Challenges
- Project Objectives
- How This Project is Structured
- ETL Strategy Used in This Project
- Designing the Data Warehouse Layers
- Flow of Tasks in Bronze Layer
- Source Systems & Business Context Questions
- Flow of Tasks in Silver and Gold Layers
- Data Flow (Lineage), Data Integration Diagram, Data Model (Star Schema)
- Data Warehouse Architecture
- SQL-Based EDA and Advanced Analytics
- GitHub Repository Structure
- Visualizing Insights in Power BI
- Key Deliverables & Future Enhancements
- Key Takeaways

Business Context & Challenges

Business Context:

- Hypothetical mid-sized retail company
- CRM System - Track sales transactions, core details of products and customers
- ERP System - Additional details of customers and products.

Challenges:

- Fragmented data across ERP and CRM systems
- Manual reporting processes with inconsistent metrics
- Inability to track customer behavior or product sales holistically

Solution:

- Building a data warehouse using the Medallion Architecture (Bronze, Silver, Gold) in SQL Server

Project Objectives

- Build a modern data warehouse using SQL Server
- Integrate ERP and CRM data into a single analytical model
- Apply cleansing, enrichment, and business rules using the Medallion Architecture
- Perform SQL-based EDA and advanced analytics
- Visualize insights using Power BI Dashboards for decision-making

How This Project is Structured

Phase 1: Building the Modern Data Warehouse (SQL Server)

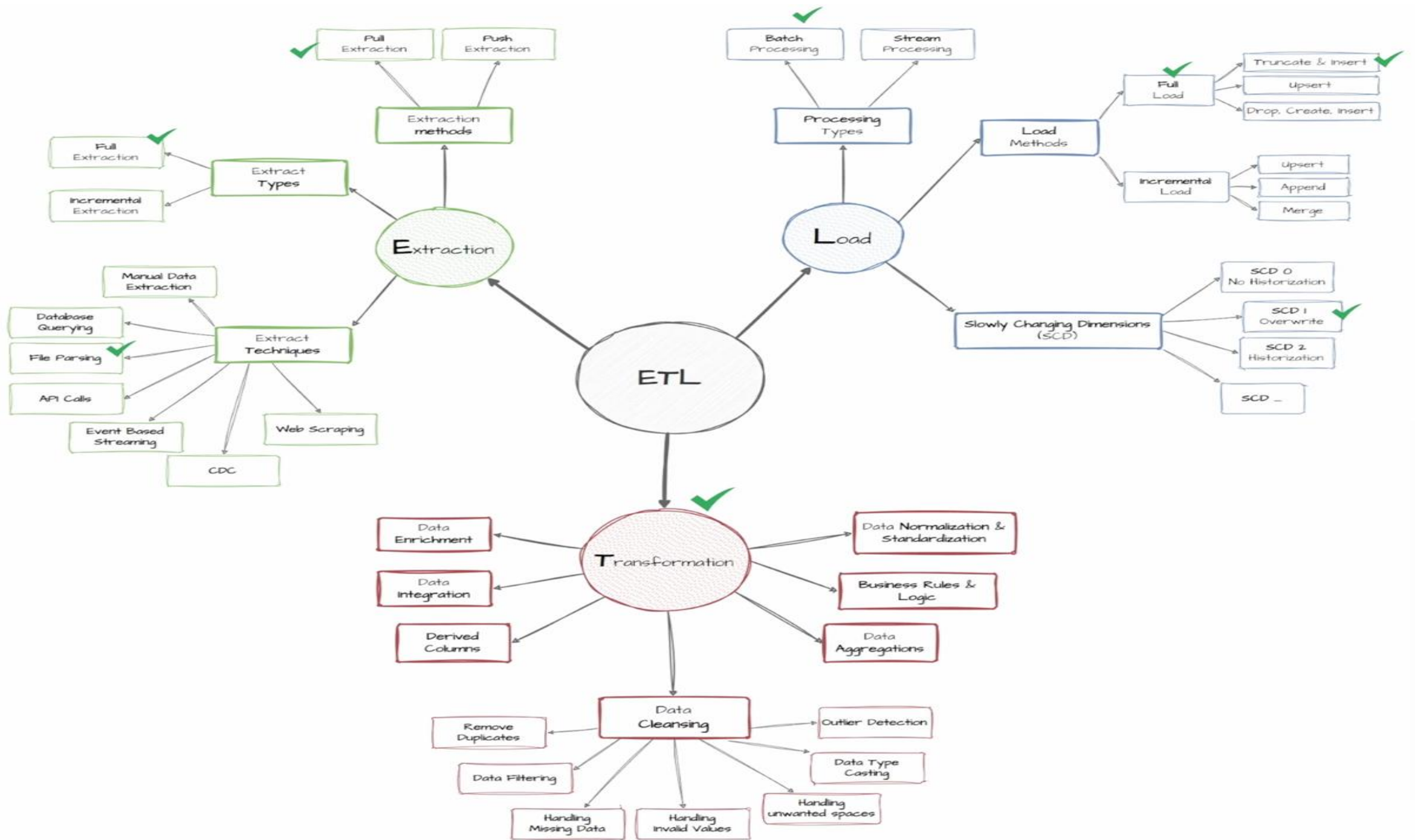
- Ingest raw ERP/CRM data
- Cleanse, enrich, and integrate using Medallion Architecture

Phase 2: EDA & Advanced Data Analytics (SQL Server)

- Perform segmentation, KPI analysis, trend analysis using SQL

Phase 3: Business Dashboards in Power BI

- Visualize insights with interactive reports and executive dashboards





Bronze Layer



Silver Layer



Gold Layer

Definition	Raw, unprocessed data as-is from sources	Clean & standardized data	Business-Ready data
Objective	Traceability & Debugging	(Intermediate Layer) Prepare Data for Analysis	Provide data to be consumed for reporting & Analytics
Object Type	Tables	Tables	Views
Load Method	Full Load (Truncate & Insert)	Full Load (Truncate & Insert)	None
Data Transformation	None (as-is)	<ul style="list-style-type: none">- Data Cleaning- Data Standardization- Data Normalization- Derived Columns- Data Enrichment	<ul style="list-style-type: none">- Data Integration- Data Aggregation- Business Logic & Rules
Data Modeling	None (as-is)	None (as-is)	<ul style="list-style-type: none">- Start Schema- Aggregated Objects- Flat Tables
Target Audience	<ul style="list-style-type: none">- Data Engineers	<ul style="list-style-type: none">- Data Analysts- Data Engineers	<ul style="list-style-type: none">- Data Analysts- Business Users



Bronze Layer

Analyzing

Interview Source Systems
Experts



Coding

Data Ingestion



Validating

Data Completeness
Schema Checks

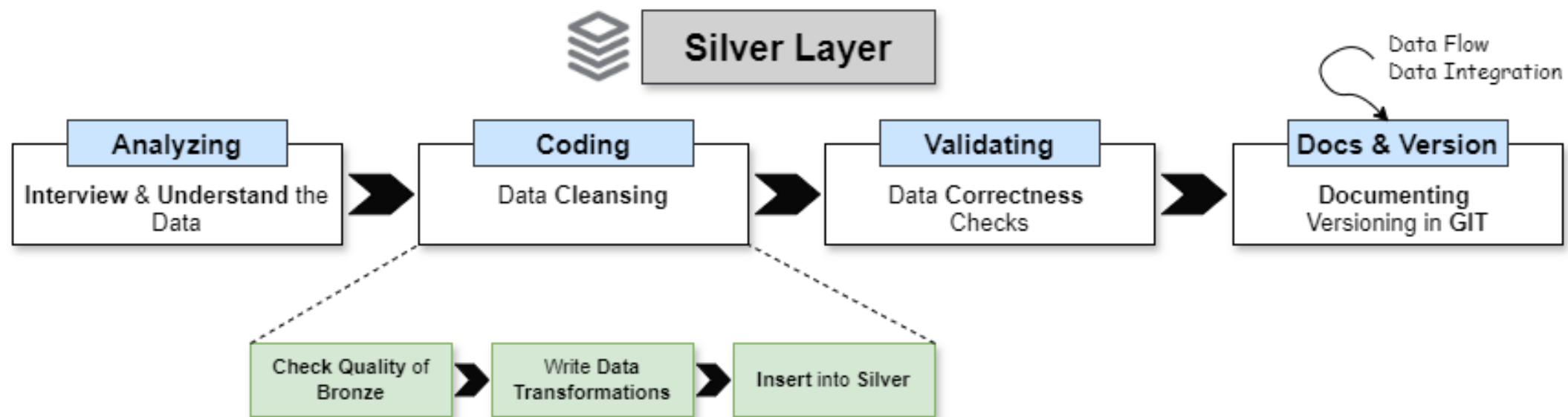


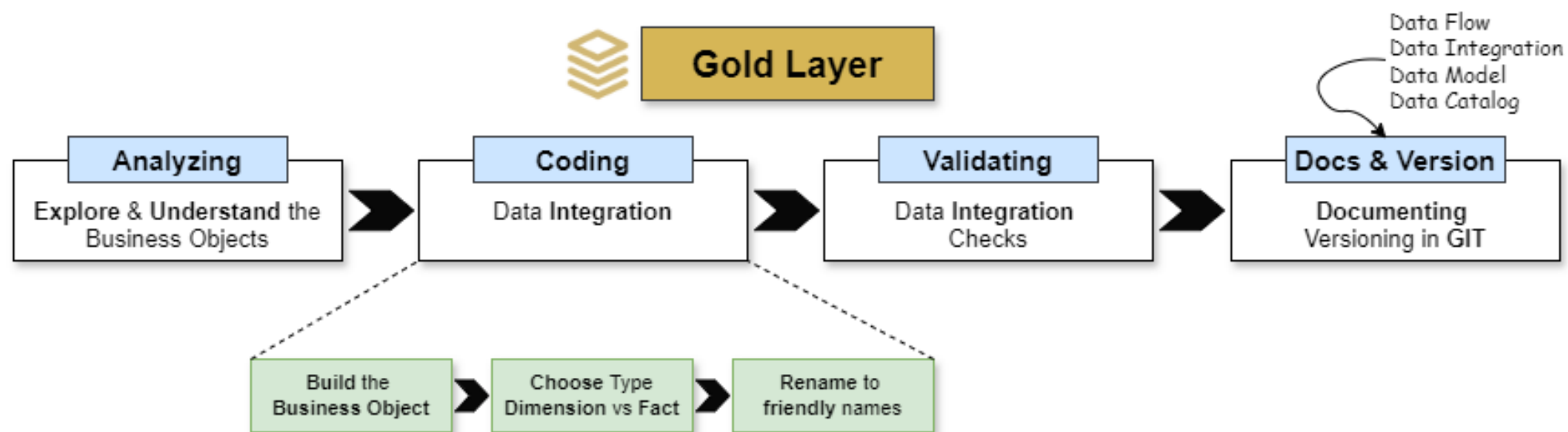
Docs & Version

Documenting
Versioning in GIT

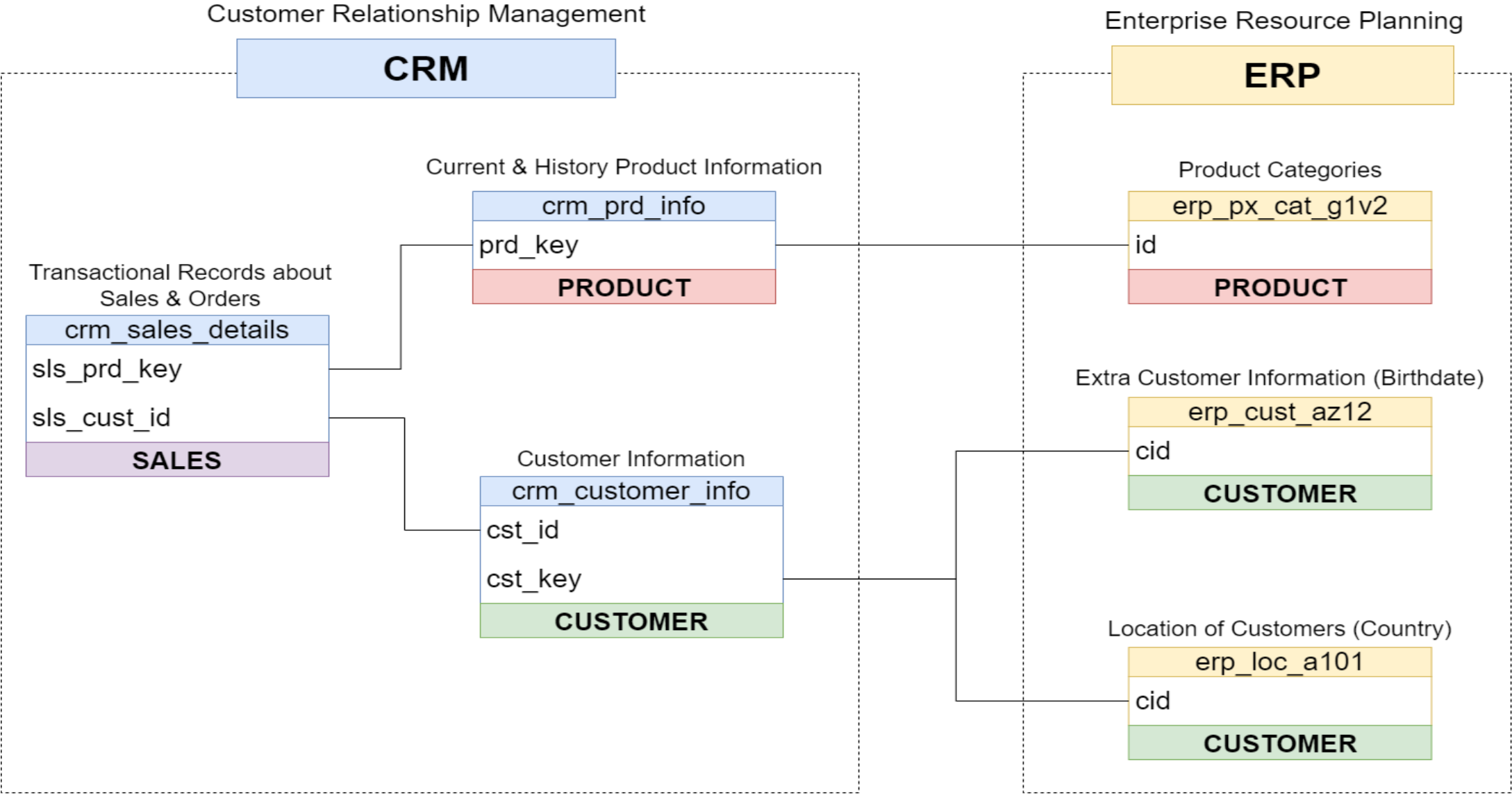
Source Systems & Business Context Questions

- Who owns the data in each system (ERP, CRM, etc.)?
- What business processes do these systems support (e.g., Sales, Customer Relationship)?
- What are the data formats and storage mechanisms (CSV, SQL Server, Oracle, cloud storage)?
- What are the integration capabilities? (API, file extracts, direct DB access, streaming tools like Kafka?)
- What is the authentication or access control mechanism? (Tokens, VPN, SSH, whitelisting?)
- What are the peak load times or usage periods for these systems?
- What is the frequency of data refresh or updates (daily, hourly, real-time)?
- Do we require full loads or incremental (delta) loads?
- How large are the typical data extracts? Are there any volume constraints?
- Are there known issues with data quality or completeness?
- What fields are most critical for business reporting and KPIs?
- How do we validate the correctness of the data post-ingestion?
- What level of historization is required (if any)?
- What are the reporting pain points that a data warehouse is expected to solve?





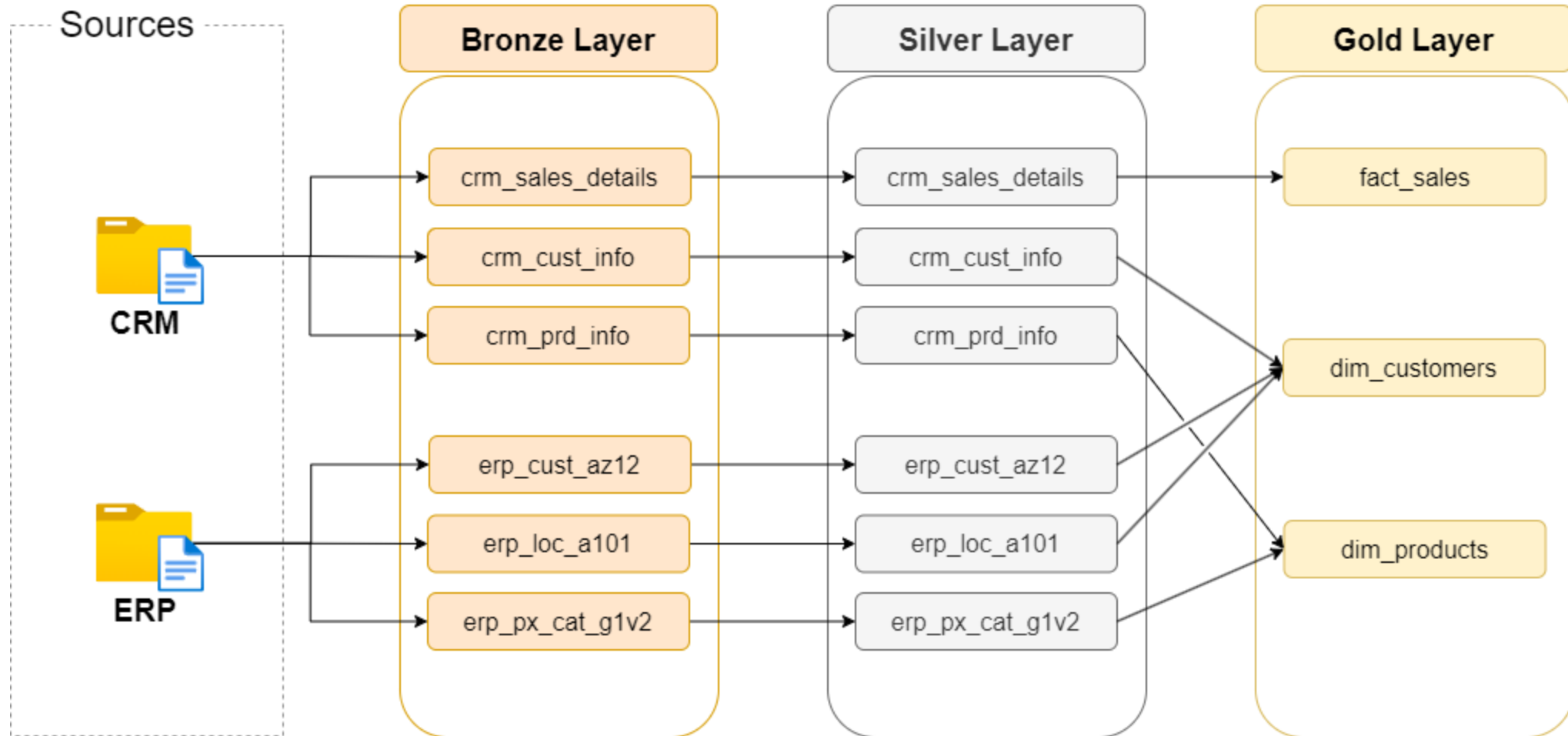
Data Integration (Source Data Model - How Tables are Connected)



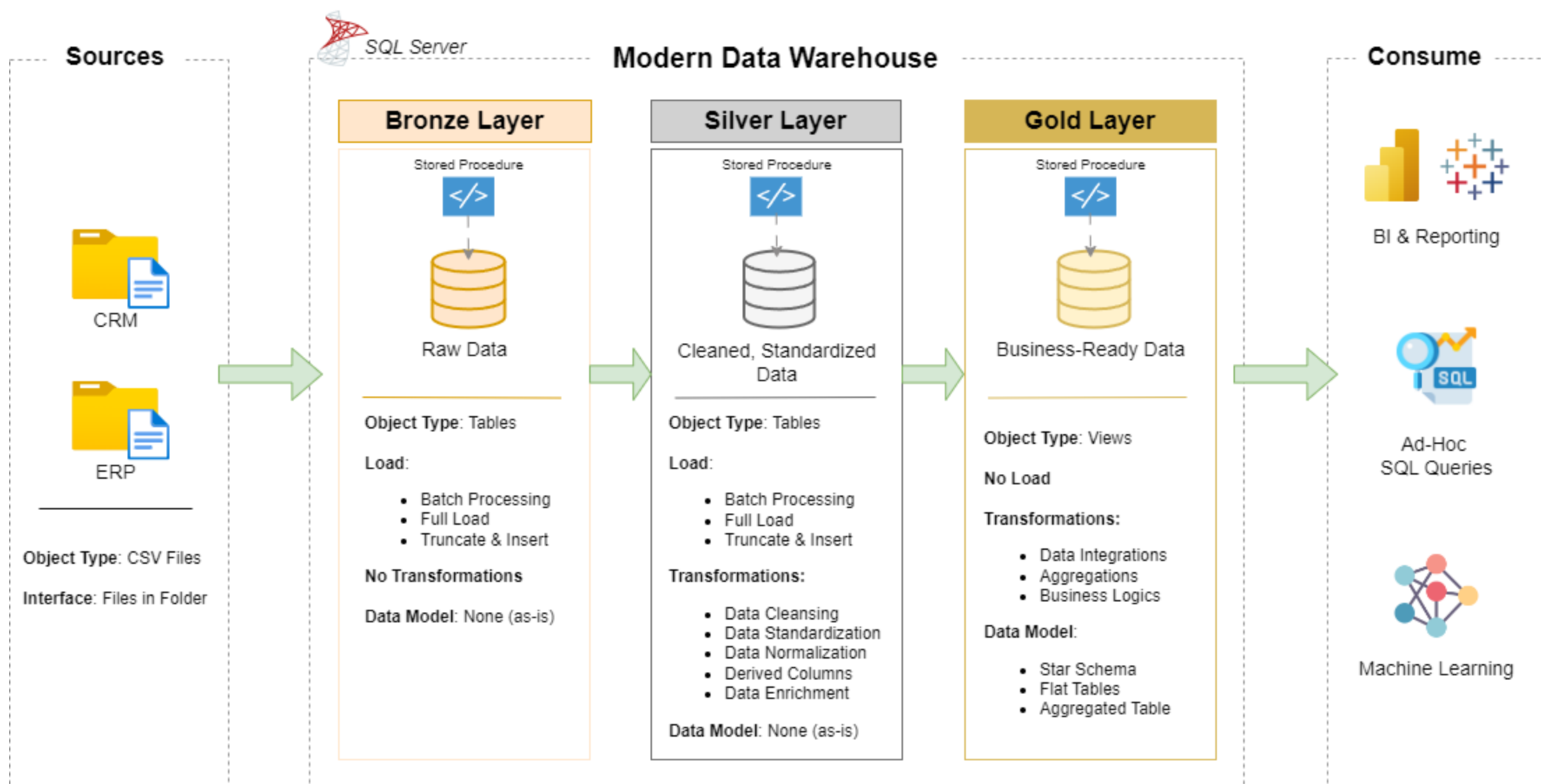
Data Model (Star Schema)



Data Flow (Lineage) Diagram



High Level Data Architecture



SQL-Based EDA and Advanced Analytics – Phase 2

Exploratory Data Analysis (EDA):

- Database Exploration
- Dimensions Exploration
- Date Exploration
- Measure Exploration
- Magnitude Analysis
- Ranking Analysis

Advanced Data Analytics:

- Change Over Time Analysis
- Cumulative Analysis
- Performance Analysis
- Part to Whole Analysis
- Data Segmentation
- Build Customers Report
- Build Products Report

data-warehouse-and-analytics-project/

```
└─ datasets/
```

```
# Raw datasets used for the project (ERP and CRM data)
```

└─ docs/

```
# Project documentation and architecture details
```

data architecture.drawio

```
# High-level project architecture (Bronze, Silver, Gold)
```

| | data catalog.md

```
# Catalog of datasets, including field descriptions and metadata
```

```
| |— data flow.drawio
```

```
# Visual representation of data flow across layers
```

```
| |— data flow tasks.drawio
```

```
# Flow of tasks in each layer – analyzing, coding, validating, documenting
```

data_integration.drawio

```
# Visual representation that depicts how Source Tables are connected
```

```
| | — data_layer_specifications.drawio
```

```
# Summarizes the objectives, transformations, and targets of each layer
```

```
| | | data_model.drawio
```

```
# Data model design (e.g., star schema)
```

etl methods.md

```
# Brief explanation of ETL strategy and methods used in this project
```

| | etl_mind_map.png

```
# Mind map showing the holistic understanding of ETL
```

```
|  └─ naming_conventions.md
```

```
# Consistent naming guidelines for tables, columns, and files
```

```
└─ scripts/
```

All SQL-based work divided into two main tracks

```
| └─ data warehouse/
```

```
# Scripts for building the data warehouse
```

```
| | | └─ bronze/
```

```
# Scripts for extracting and loading (full load) raw data
```

```
| | | └─ silver/
```

```
# Scripts for cleaning and transforming data
```

```
| | | | | L gold/
```

```
# Scripts for creating analytical models (views and data models)
```

```
|  └─ eda_analytics/
```

```
# Scripts for FDA and advanced data (business) analytics
```

└─ tests/

```
# QA scripts for verifying integrity and logic of gold and silver layers
```

 README.md

Project overview and instructions



REVENUE

\$16.3M✓

Prev Year: \$5.8M (+179.77%)

PROFIT

\$6.8M✓

Prev Year: \$2.1M (+228.83%)

ORDERS

21287✓

Prev Year: 3269 (+551.18%)

AVERAGE ORDER VALUE

\$767.8!

Prev Year: \$1.8K (-57.04%)

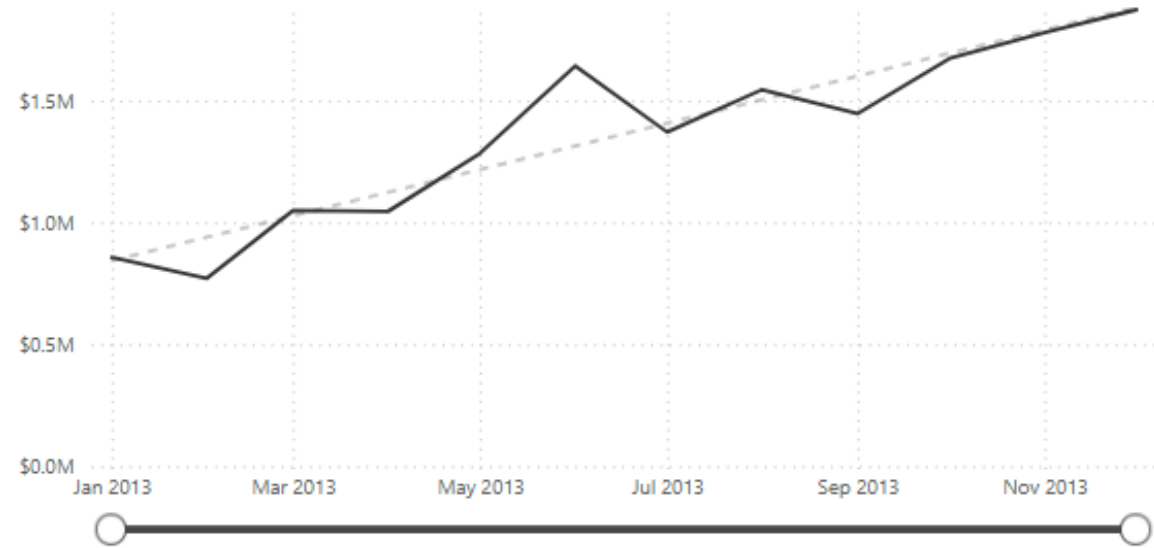
TopNValue

30

Top N Filter By

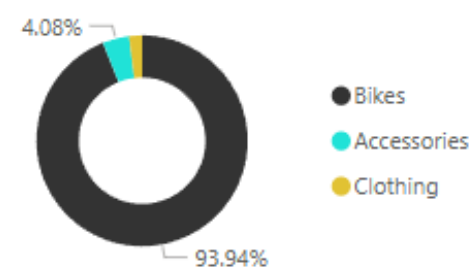
- ☐ Avg Order Value
- ☐ Orders
- ☒ Revenue

Revenue Trending

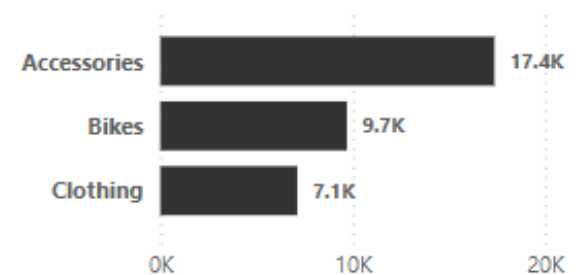


Sub-Category	Revenue	Orders	Avg Order Value
Mountain Bikes			
Mountain-200 Black- 42	\$970,785	423	\$2,295
Mountain-200 Silver- 38	\$967,440	417	\$2,320
Mountain-200 Black- 46	\$947,835	413	\$2,295
Mountain-200 Black- 38	\$945,540	412	\$2,295
Mountain-200 Silver- 46	\$921,040	397	\$2,320
Mountain-200 Silver- 42	\$902,480	389	\$2,320
Road Bikes			
Road-350-W Yellow- 40	\$416,745	245	\$1,701
Road-350-W Yellow- 42	\$394,632	232	\$1,701
Road-350-W Yellow- 48	\$392,931	231	\$1,701
Road-250 Black- 48	\$381,108	156	\$2,443
Road-350-W Yellow- 44	\$367,416	216	\$1,701
Road-250 Black- 52	\$351,792	144	\$2,443
Road-250 Black- 44	\$342,020	140	\$2,443
Road-250 Red- 58	\$320,033	131	\$2,443
Road-250 Black- 58	\$302,932	124	\$2,443
Road-550-W Yellow- 42	\$265,440	237	\$1,120
Road-550-W Yellow- 44	\$234,080	209	\$1,120
Road-550-W Yellow- 38	\$221,760	198	\$1,120

Revenue by category



Orders by category



Key Deliverables & Future Enhancements

Key Deliverables:

- A fully functional, SQL Server-based data warehouse, based on Medallion Architecture
- Clean and reusable SQL scripts for analytics
- Documentation for the data model, SQL Scripts, architecture diagrams, and so on.
- Power BI Dashboards, shared online through Power BI Service

Future Enhancements:

- Automating incremental loads via Change Data Capture (CDC)
- Introducing historical change tracking via SCD Type 2 for slowly changing dimensions (e.g., customer segment or product category shifts)

Key Takeaways

- Simulated a realistic end-to-end data project — from raw ERP/CRM data to insights-ready dashboards
- Strengthened SQL, data modeling, and analytics problem-solving through layered architecture
- Practiced stakeholder-style thinking — aligning KPIs with actual business questions and decisions
- Gained hands-on experience with project structuring, documentation, and insight storytelling