Mercedes-Benz

Regular Day & Main Tasks of a Data Engineer

A regular day for a Data Engineer in an Agile team might look like this:

Morning Stand-Up: Discuss with the Team, done, to do and blocking points.

Development Work: Focusing on ETL processes, building, testing, and optimizing data transformations, and data pipeline improvements.

Monitoring Existing Streams: assure completion and high performance standards of existing pipeline, troubleshoot and resolve bugs and issues.

Collaboration: Working with Platforms, BI and Analytics to ensure data quality and accessibility.

Code Review & Testing: Ensuring ETL processes adhere to business rules and data quality standards.

Main Tasks and Responsibilities

Design, construct, install, and maintain large-scale processing systems and other infrastructure.

Ensure Systems Meet Business Requirements and Industry Practices

Build high-performance processes, prototypes, and proof of concepts.

Research opportunities for data acquisition and new uses for existing data.

Preparation for Design Session on New Star-Schema Feature:

Requirement Gathering: Understand all the requirements and constraints around the survey data.

Explore Data: Initial assessment of survey data structure in Salesforce.

API Exploration and Testing: Engage with the Salesforce Developer to comprehend how to access and test the Salesforce API. Consider using tools like Postman for preliminary testing.

Destination Source Setup: Engage with DBAs to set up the destination source, taking into consideration the size needed and the security levels.

Data Mart Decisions: Determine whether to create a new Data Mart or integrate this into the existing Cases Data Mart, especially since it has a link to it. This evaluation will further drive the decision on whether this is a new ETL project or an extension of the Cases ETL.

Collaboration with BI and Analytics Teams: Engage with BI and Analytics teams to shape the ETL in a manner that outputs data ready for BI, reducing the need for further transformations on their end.

Business Logic Considerations: If metrics like the NPS aren't automatically calculated by Salesforce, the ETL should handle such calculations. This ensures that BI and Analytics won't need to embed additional logic in their tools.

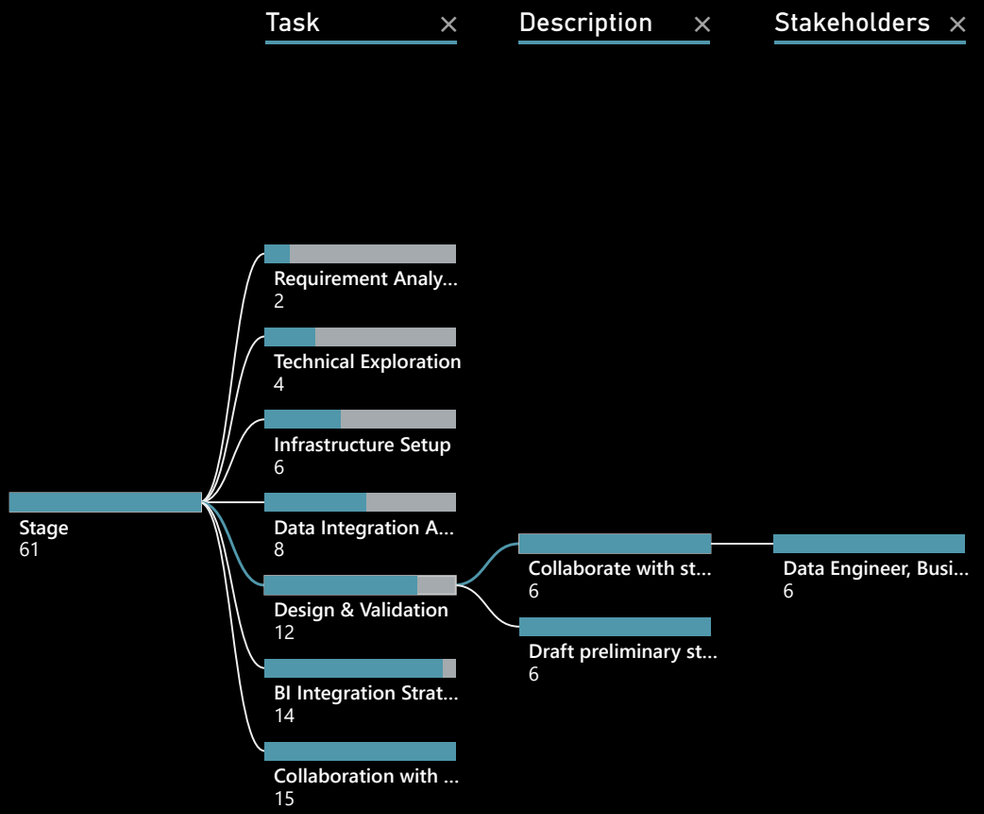
Participation in Business Requirements: Actively take part in business requirements gathering sessions to gain a deeper understanding of data usage, ensuring the data transformation aligns with business needs.

Table Referencing with BI: Coordinate with the BI team to determine optimal ways to reference tables directly in the BI model, minimizing the need for hardcoded values in BI tools and ensuring that data remains clear and centralized in the DWH.

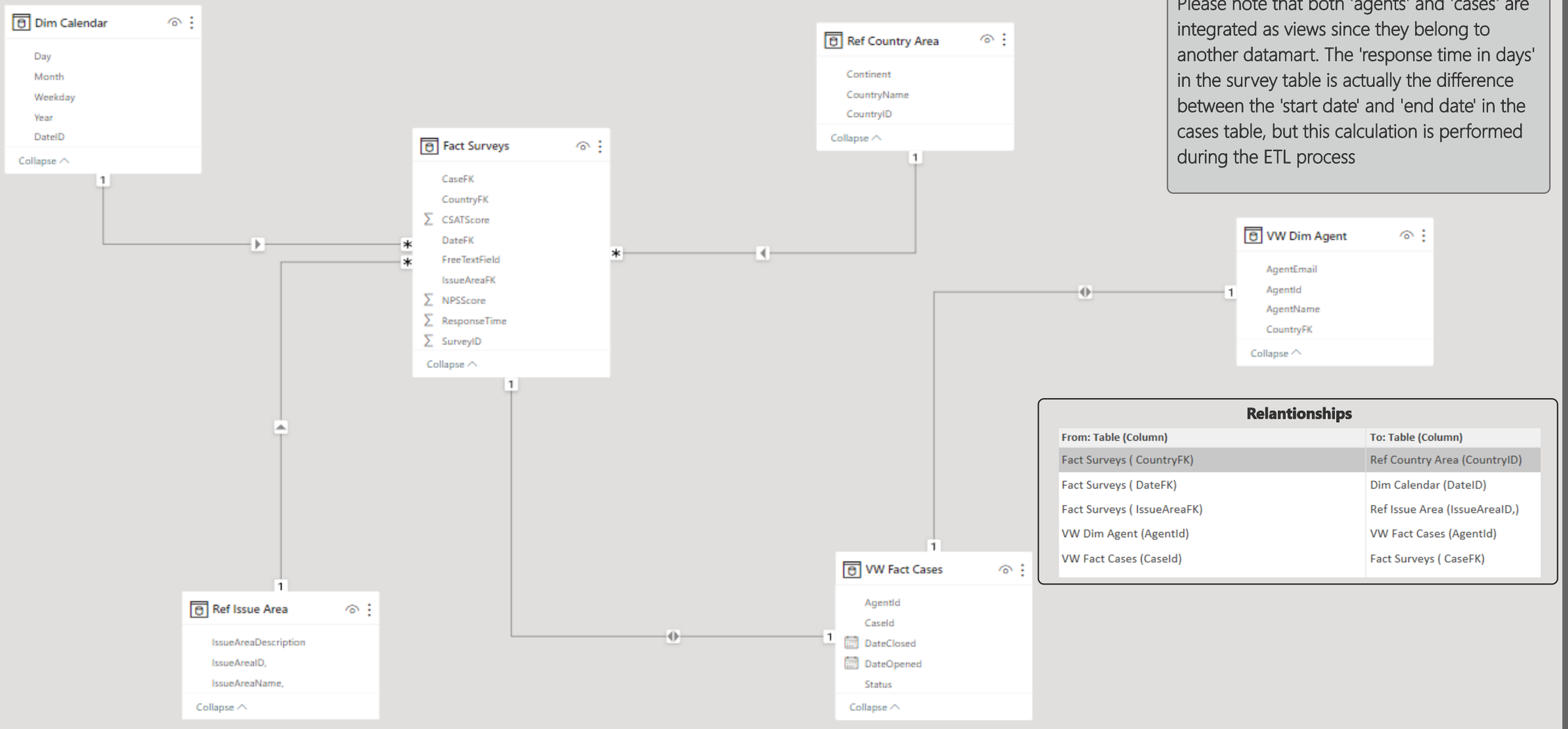
Preliminary Design: Sketch a high-level design of the star-schema incorporating CSAT score, NPS score, response time, and other metrics.

Collaboration: Discuss with data architects, Salesforce experts, and business analysts to validate preliminary design and requirements.

Workflow Process



Data Model



Please note that both 'agents' and 'cases' are integrated as views since they belong to another datamart. The 'response time in days' in the survey table is actually the difference between the 'start date' and 'end date' in the cases table, but this calculation is performed during the ETL process

Relantionships

From: Table (Column)	To: Table (Column)
Fact Surveys (CountryFK)	Ref Country Area (CountryID)
Fact Surveys (DateFK)	Dim Calendar (DateID)
Fact Surveys (IssueAreaFK)	Ref Issue Area (IssueAreaID,)
VW Dim Agent (AgentId)	VW Fact Cases (AgentId)
VW Fact Cases (CaseId)	Fact Surveys (CaseFK)

Database Objects

Mercedes-Benz

Fact Surveys

SurveyID	CaseFK	DateFK	IssueAreaFK	CountryFK	CSATScore	NPSScore	ResponseTime	FreeTextField
1	1001	20230101	1	1	4	10.00	10	Helpful sales representative!
2	1002	20230102	2	2	5	10.00	12	Quick mechanical repair!
4	1004	20230103	4	11	4	10.00	8	Good Service, Loan was quick
5	1005	20230104	5	2	4	10.00	9	The agent was kind and quick to resolve the issue
6	1006	20230105	1	10	5	10.00	7	Sales team provided great financing options.
9	1009	20230108	4	5	5	10.00	6	Top Tier Customer Service! 5 Stars!!!!!
10	1010	20230109	5	5	5	10.00	8	Fast, Easy, Efficient! Mercedes is the best!
3	1003	20230102	3	1	3	5.00	15	Software update was confusing.
7	1007	20230106	2	3	3	5.00	16	The mechanical noise persists.
8	1008	20230107	3	1	2	0.00	18	Software lacks intuitive features.

VW Fact Cases

CaseId	Agen...	DateOpened	DateClosed	Status
1001	A001	1/1/2023	1/11/2023	Closed
1002	A002	12/21/2022	1/2/2023	Closed
1003	A003	12/18/2022	1/2/2023	Closed
1004	A004	12/26/2022	1/3/2023	Closed
1005	A005	12/26/2022	1/4/2023	Closed
1006	A006	12/29/2022	1/5/2023	Closed
1007	A007	12/21/2022	1/6/2023	Closed
1008	A008	12/20/2022	1/7/2023	Closed
1009	A009	12/23/2022	1/8/2023	Closed
1010	A010	12/31/2022	1/9/2023	Closed

Ref Country

CountryID	CountryName	Continent
1	USA	North America
2	Canada	North America
3	Germany	Europe
4	UK	Europe
5	Netherlands	Europe
6	Italy	Europe
7	Greece	Europe
8	Brazil	South America
9	Japan	Asia
10	United Arab Emirates	Asia
11	Saudi Arabia	Asia

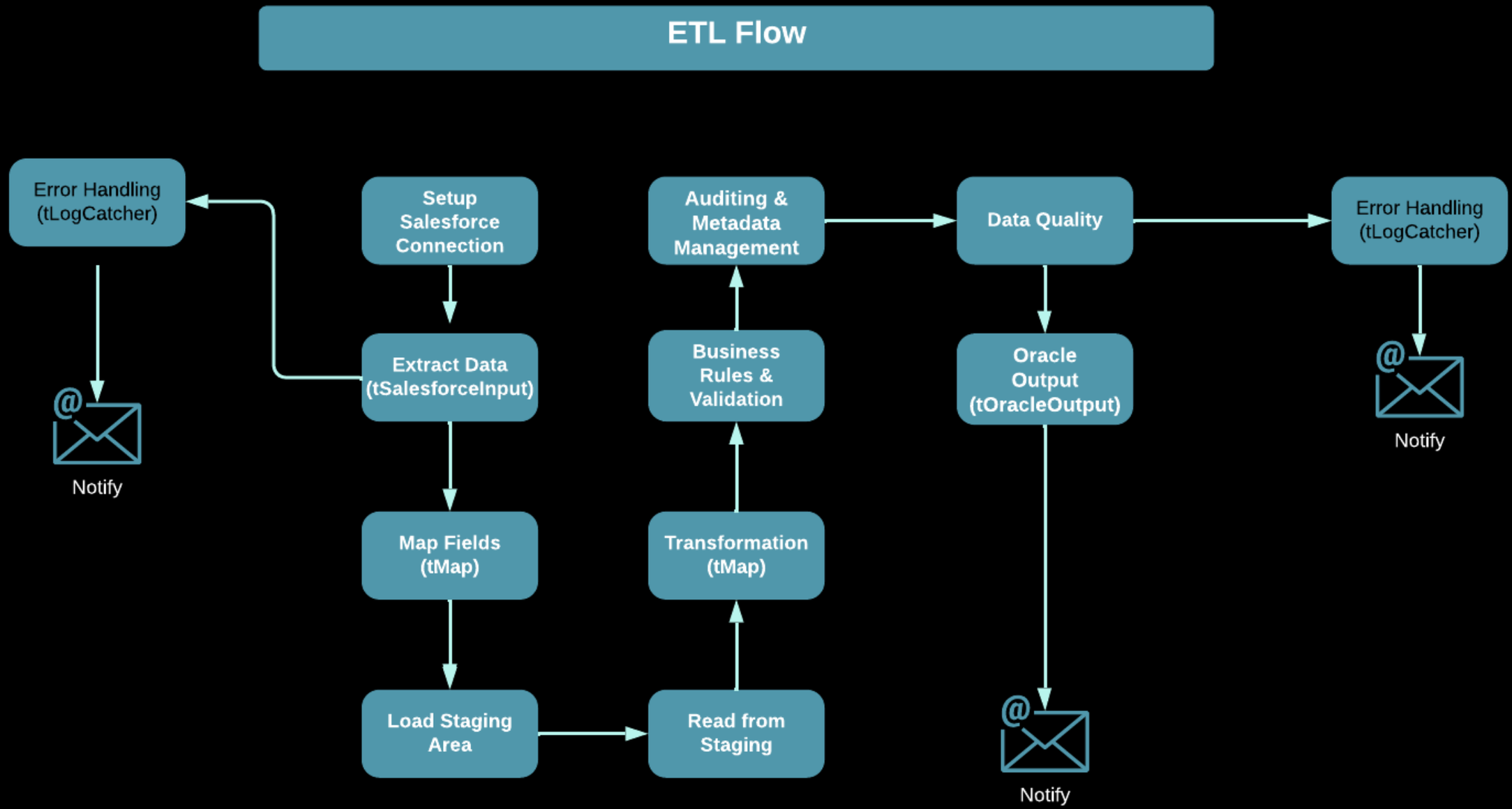
VW Dim Agent

AgentId	AgentName	AgentEmail	CountryFK
A001	Johannes de Vries	johannesdevries@CACMercedesbenz.com	5
A002	Eleni Papadopoulos	elenipapadopoulos@CACMercedesbenz.com	7
A003	Bram van der Meer	bramvandermeer@CACMercedesbenz.com	5
A004	Nikos Vasilis	nikosvasilis@CACMercedesbenz.com	7
A005	Sofie Jansen	sofiejansen@CACMercedesbenz.com	7
A006	Maria Stavros	mariastavros@CACMercedesbenz.com	7
A007	Pieter de Jong	pieterdejong@CACMercedesbenz.com	5
A008	Dimitris Ioannou	dimitrisioannou@CACMercedesbenz.com	7
A009	Fleur van den Berg	fleurvandenberg@CACMercedesbenz.com	5
A010	Katerina Georgiou	katerinageorgiou@CACMercedesbenz.com	7

Ref Issue Area

IssueAreaID	IssueAreaName	IssueAreaDescription
1	Retail Sales	Feedback related to vehicle sales, offers, and purchasing.
2	Mechanical	Concerns regarding vehicle mechanics and maintenance.
3	Software	Feedback on onboard software, updates and navigation features.
4	Finance	Feedback on financing and Loans assistance.
5	Corporate Sales	Experiences with customer service representatives and assistance.

ETL Flow


ETL Flow
Error Handling (tLogCatcher)
Notify
Setup Salesforce Connection
Extract Data (tSalesforceInput)
Map Fields (tMap)
Load Staging Area
Read from Staging
Transformation (tMap)
Business Rules & Validation
Auditing & Metadata Management
Data Quality
Oracle Output (tOracleOutput)
Notify
Error Handling (tLogCatcher)
Notify

Development Jira Roadmap

JiraS1

Title: Extract Survey Data from Salesforce

Task: Collaborate with Salesforce admin to set up appropriate API endpoints. Ensure data security by encrypting sensitive data during extraction. Schedule regular extraction jobs and handle potential API rate limits.

Role: Data Engineer \ Salesforce Adminstrator

JiraS2

Title: Transform Data for and Optimize for BI (Calculation of basic KPI's)

Task: - Load raw data into a staging area, preserving original source integrity. - Convert staged data to the star-schema structure, ensuring DWH schema alignment. - Address data quality: handle duplicates, missing values, and inconsistencies. - Calculate KPIs such as NPS from CSAT, response times, SLA checks, and elapsed times in ETL. - Implement logic in ETL to reduce post-processing and hardcoded logic in BI tools. - Flag SLA breaches and generate error logs for discrepancies during transformations.

Role: Data Engineer

JiraS3

Title: Load Transformed Data into DWH

Task: Design efficient loading processes to handle large data volumes. Ensure data consistency and integrity with validation checks. Optimize for incremental data loads to reduce load times and system impact. Testing with the BI Team

Role: Data Engineer

JiraS4

Title: Build and Optimize Data Model in PowerBI

Task: Collaborate with business stakeholders to understand reporting needs. Design efficient data models and relationships in PowerBI. Optimize reports for speed and user experience. Provide training sessions for business users on how to use the new reports.

Role: BI Developer

JiraS5

Title: Develop Text Mining Model for Survey Feedback

Task: Analyze free text fields to identify common themes and sentiments. Collaborate with business analysts to understand the type of insights they are looking for and refine the model accordingly.

Role: Data Scientist

Mercedes-Benz

In the 3rd sprint of the PI, you get an urgent defect from production assigned. Seems there is an issue with duplicate data records that we loaded in Salesforce production. This will impact your progress on the committed feature. After analyzing the defect for a few hours, you realize you will need a lot more time to find the root cause and fix the defect. The time reserved for defect handling this PI is already fully consumed by the agile team.

How will you handle this situation?

- **Notify the Team Lead\ Scrum Master:** As soon as the defect is understood, promptly notify the Scrum Master, Product Owner, and Team Lead. Transparent communication is crucial when we're nearing the end of the PI and facing potential disruptions.
- **Assess Impact and Severity:** Jointly with the Team Lead, understand the severity of the issue. Discuss its ramifications on business operations and how stakeholders might be affected. If the defect has severe implications, consider communicating to the broader business the potential delays.
- **PI Go Live Date:** With the defect occurring in the 3th sprint of the PI, review the committed goals with the Product Owner. Recognize that the defect's resolution might cause delays, potentially shifting the go-live date.
- **Rescheduling and Realignment:** If it's determined that go-live may be delayed, devise a plan with the team to reallocate resources effectively. This may mean re-prioritizing tasks, organizing additional development and testing sessions, or even extending the sprint, if possible.

Which technical approaches would you use to identify the duplicate records?

- **Database Check:** Utilize SQL tools identify repetitive data based on certain criteria.
- **Check with source owner:** Before diving deep into debugging, connect with the source data owner to understand if the root of the issue lies with the source data or elsewhere in the process flow. Establish if the issue stems from user **data entry mistakes** or system-induced
- **Investigate Data Pipeline:** After assessing that the source is correct, investigate errors in the ETL maybe caused by wrong joins in algorithms or duplicate loads. Check logs to understand if certain data was loaded twice in a single batch window
- **Create Steps to avoid future disruptions**: Implement steps like ETL Data Quality Checks, Table Constraints & Lookups and engage with the PO to understand if certain constraints, that usually are custom ones, are implemented in the product.

NPSScore
8.00

Average of ResponseTime
10.90

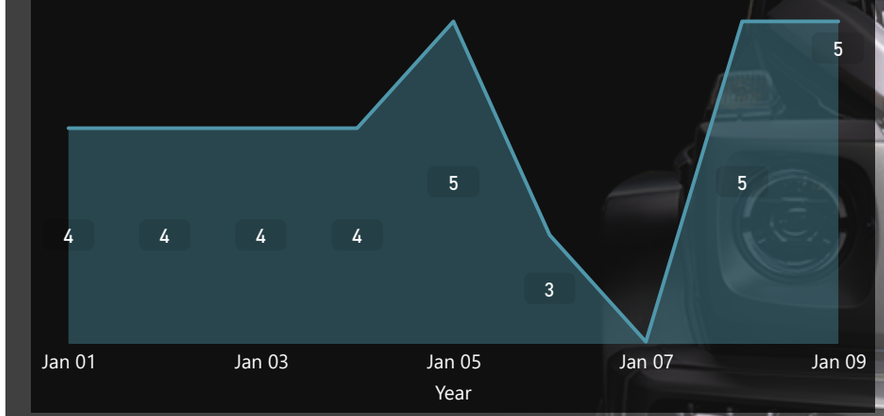
% Detractors
16.67%

% Promoters
50.00%

Top AVG CSAT Area
Corporate Sales

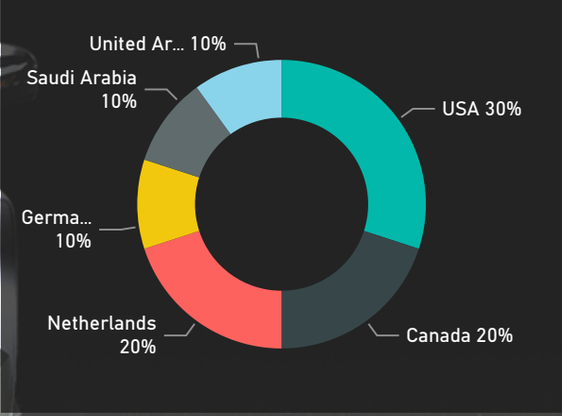
Top AVG CSAT Country
Brazil

Average CSAT Score By Time

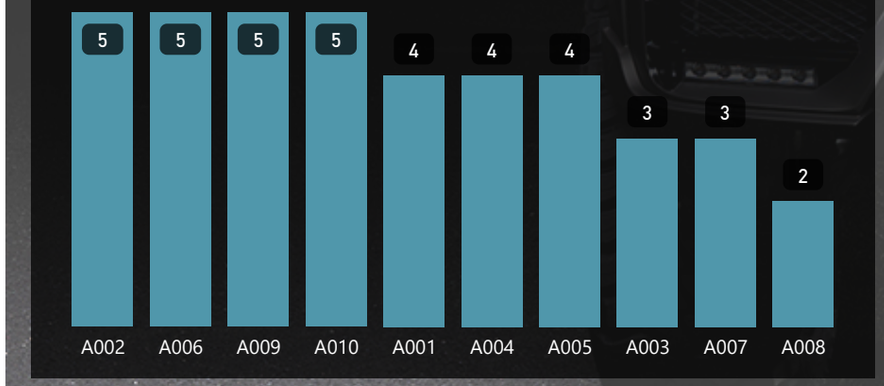


Average CSAT per Area

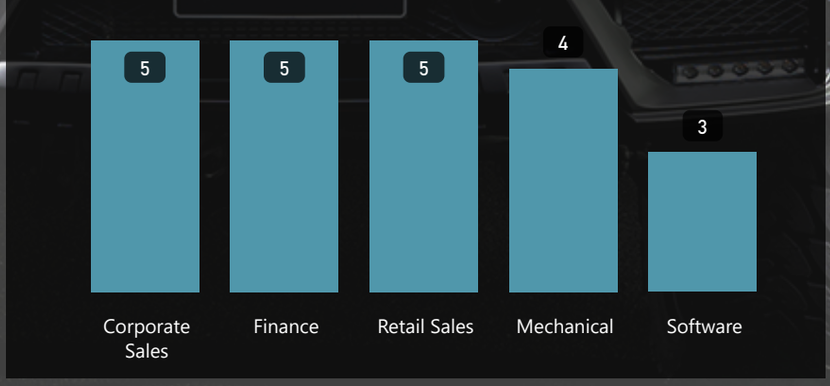
Cases per Country



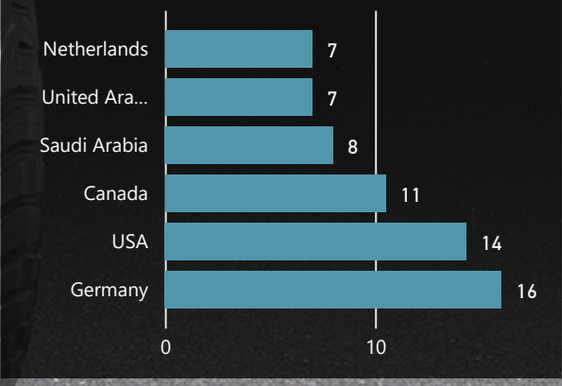
Average CSAT per Agent



Average CSAT per Area



VAG Response Time per Country



Mercedes-Benz

Conclusions:

- **Integrated Tables - Cases:** I've utilized a materialized view based on the 'case facts' table, focusing exclusively on closed cases. This approach is based on my assumption that surveys are conducted only after a case has been resolved.
- **Integrated Table - Agents:** For metrics related to agents and insights, I sourced data from the 'agent view', which provides in-depth agent information.
- **Audit Fields:** To ensure transparency and traceability in my data, I would have audit fields. These fields capture the 'last load date' and the 'ETL project name', but they won't be visible in the BI layer.
- **Logging:** I've designed the ETL process to leverage an already established logging engine, ensuring that all logs from this process are integrated into existing tables seamlessly.
- **Monitoring Dashboard:** If there's an existing monitoring dashboard, my new process will automatically be integrated, streamlining the sanity checks for all ETLs.
- **Historical Tables:** Given that the data pertained to Events and Slowly Changing Dimensions (SCD), I opted not to implement an intricate historical tracking process. If required, I would have taken advantage of integrated Temporal Tables in SQL Server or Oracle to manage data versioning. In other scenarios, particularly when there's no need to modify past events, each event can be timestamped. The associated dimension can be equipped with start and end dates, along with an 'active' flag. This facilitates filtering during queries, ensuring that OLAP cubes consider only the currently active records.

Mercedes-Benz

Thanks for the attention