mETal

Abrosov, Anisin, Lanin, Malysh, Pan

Project statement

Ensuring high data quality and availability is crucial for applications operating within our framework. Our extension for VS Code utilizes “data modules,” which are stored either on Amazon S3 or in a relational database management system (RDBMS). The ETL (Extract, Transform, Load) service is designed to pull data from internal corporate systems, perform extensive cleansing, merging, anonymization, and transformation of the data, and subsequently upload it to create a new version of the data module.

Team: Micro SD

Members:

- Abrosov Sergey
- Anisin Aleksandr
- Lanin Georg
- Malysh Igor
- Pan Zhengwu

Project repo: https://github.com/abrosov-sergey/Micro-SD.git

This report: https://clck.ru/3DDkiu

Technical roles

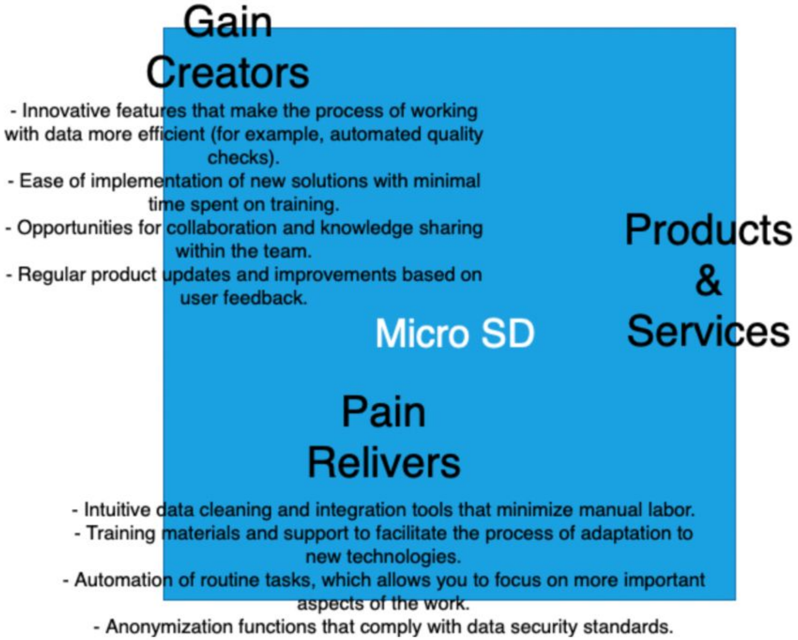
Data Engineer

Our service can be useful for data engineers. As data engineers, users can create high-quality, anonymized datasets on demand. So they can clean, filter, merge, anonymize and export data from the company's information systems.

MLOps Engineer

It can also be useful for MLOps engineers. The service can help with regular support of ML pipelines. It will provide data for the selected ML models automatically.

Value proposition canvas



Gain Creators

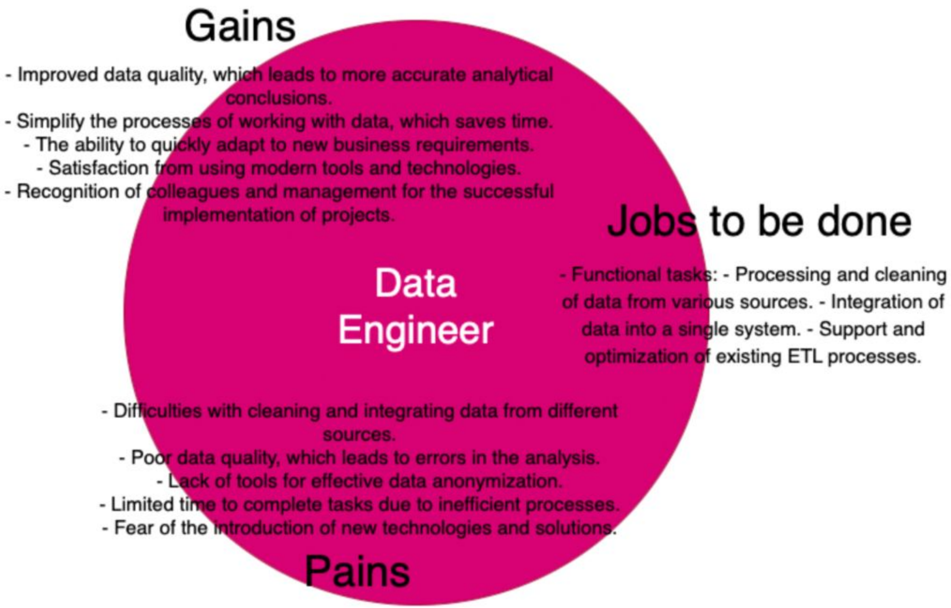
- Innovative features that make the process of working with data more efficient (for example, automated quality checks).
- Ease of implementation of new solutions with minimal time spent on training.
- Opportunities for collaboration and knowledge sharing within the team.
- Regular product updates and improvements based on user feedback.

Micro SD

Products & Services

Pain Relivers

- Intuitive data cleaning and integration tools that minimize manual labor.
- Training materials and support to facilitate the process of adaptation to new technologies.
- Automation of routine tasks, which allows you to focus on more important aspects of the work.
- Anonymization functions that comply with data security standards.



Gains

- Improved data quality, which leads to more accurate analytical conclusions.
- Simplify the processes of working with data, which saves time.
- The ability to quickly adapt to new business requirements.
- Satisfaction from using modern tools and technologies.
- Recognition of colleagues and management for the successful implementation of projects.

Data Engineer

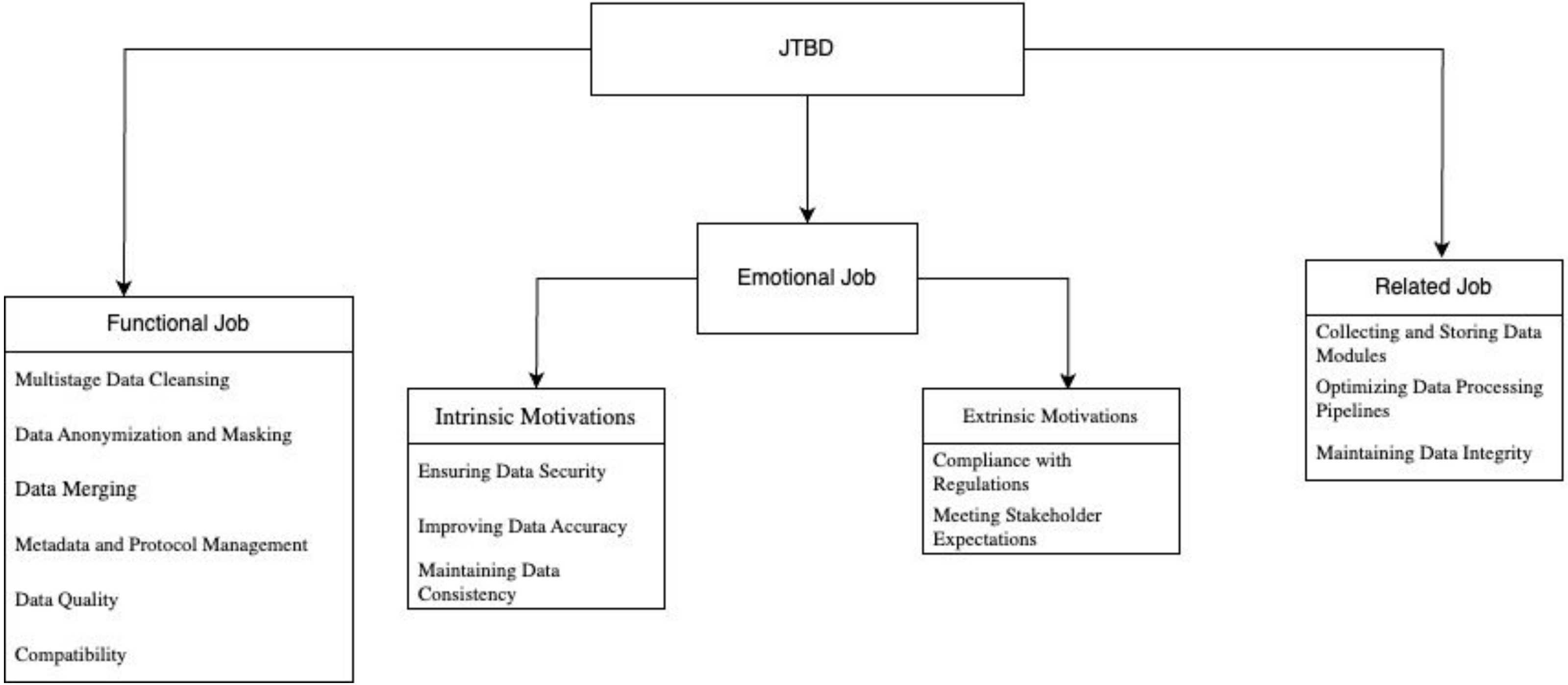
Jobs to be done

- Functional tasks: - Processing and cleaning of data from various sources. - Integration of data into a single system. - Support and optimization of existing ETL processes.

- Difficulties with cleaning and integrating data from different sources.
- Poor data quality, which leads to errors in the analysis.
- Lack of tools for effective data anonymization.
- Limited time to complete tasks due to inefficient processes.
- Fear of the introduction of new technologies and solutions.

Pains

Jobs map



JTBD

Functional Job

- Multistage Data Cleansing
- Data Anonymization and Masking
- Data Merging
- Metadata and Protocol Management
- Data Quality
- Compatibility

Emotional Job

Intrinsic Motivations

- Ensuring Data Security
- Improving Data Accuracy
- Maintaining Data Consistency

Extrinsic Motivations

- Compliance with Regulations
- Meeting Stakeholder Expectations

Related Job

- Collecting and Storing Data Modules
- Optimizing Data Processing Pipelines
- Maintaining Data Integrity

Data Glossary

Read full version in our Weeek

https://app.weeek.net/ws/633468/document/15

ETL (Extract, Transform, Load) Service

A system that extracts data from internal corporate sources, transforms it through cleansing and other processes, and loads it into data modules.

VS Code Extension

An add-on for Visual Studio Code that extends its functionality; in this project, the ETL service operates as such an extension.

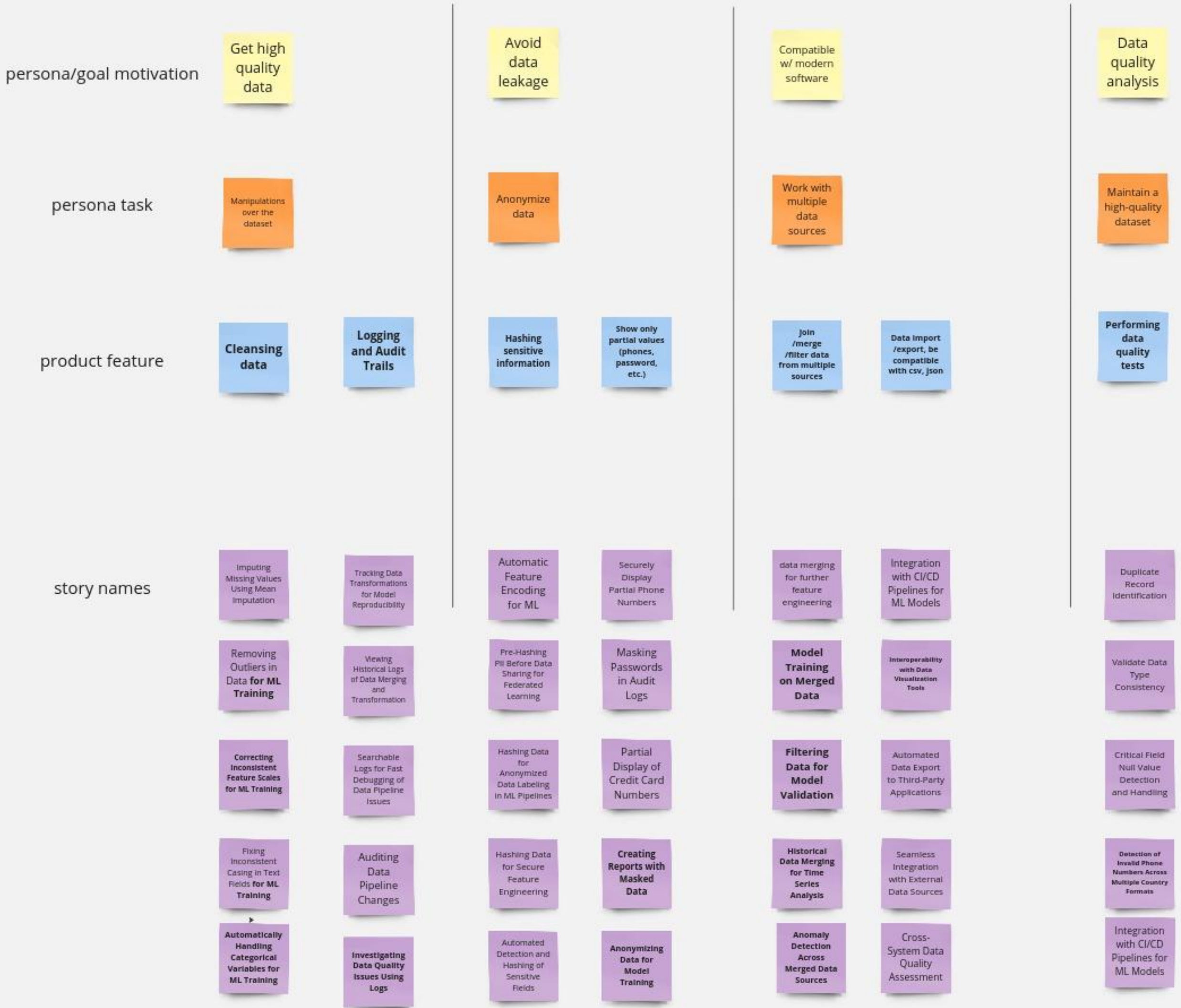
Anonymization

The process of removing or altering personally identifiable information from data to protect individual privacy while maintaining data utility for analysis.

Data Transformation

Modifying data's format, structure, or values to meet specific requirements or prepare it for further analysis.

Story map



persona/goal motivation	Get high quality data		Avoid data leakage		Compatible w/ modern software		Data quality analysis
persona task	Manipulations over the dataset		Anonymize data		Work with multiple data sources		Maintain a high-quality dataset
product feature	**Cleansing data**	**Logging and Audit Trails**	**Hashing sensitive information**	**Show only partial values (phones, password, etc.)**	**Join /merge /filter data from multiple sources**	**Data import /export, be compatible with csv, json**	**Performing data quality tests**
story names	Imputing Missing Values Using Mean Imputation	Tracking Data Transformations for Model Reproducibility	Automatic Feature Encoding for ML	Securely Display Partial Phone Numbers	data merging for further feature engineering	Integration with CI/CD Pipelines for ML Models	Duplicate Record Identification
	Removing Outliers in Data **for ML Training**	Viewing Historical Logs of Data Merging and Transformation	Pre-Hashing PII Before Data Sharing for Federated Learning	Masking Passwords in Audit Logs	**Model Training on Merged Data**	**Interoperability with Data Visualization Tools**	Validate Data Type Consistency
	Correcting Inconsistent Feature Scales for ML Training	Searchable Logs for Fast Debugging of Data Pipeline Issues	Hashing Data for Anonymized Data Labeling in ML Pipelines	Partial Display of Credit Card Numbers	**Filtering Data for Model Validation**	Automated Data Export to Third-Party Applications	Critical Field Null Value Detection and Handling
	Fixing Inconsistent Casing in Text Fields **for ML Training**	Auditing Data Pipeline Changes	Hashing Data for Secure Feature Engineering	**Creating Reports with Masked Data**	**Historical Data Merging for Time Series Analysis**	Seamless Integration with External Data Sources	**Detection of Invalid Phone Numbers Across Multiple Country Formats**
	Automatically Handling Categorical Variables for ML Training	**Investigating Data Quality Issues Using Logs**	Automated Detection and Hashing of Sensitive Fields	**Anonymizing Data for Model Training**	**Anomaly Detection Across Merged Data Sources**	Cross-System Data Quality Assessment	Integration with CI/CD Pipelines for ML Models

Story map

https://app.weeek.net/ws/633468/document/18