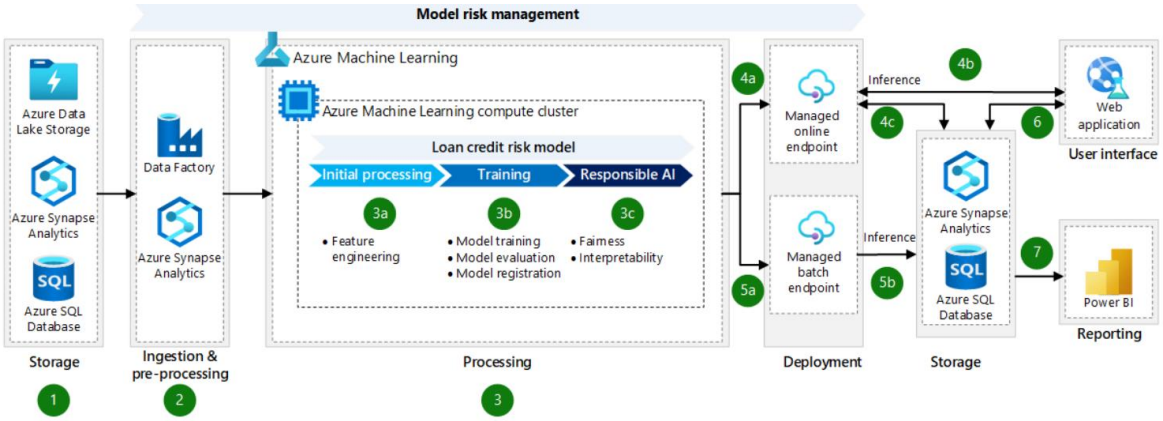
SURAJIT SHOME
AZURE DATA ENGINEER
Phone: +31 647988248

2. Define technical specifications and set of technologies that would address the following aspects:

2.1. Extract data from different data sources, Transform it to the underlying data model according to some pre-define business rules in a file format of your choice, then Load it.

2.2. Deal with structured, semi-structure and unstructured input data sources.

2.3. Scalable app in order to be able to process high-volumes of data in batch and real-time.



Data Flow

1. **Storage**: Data is stored in a database like an Azure Synapse Analytics pool if it's structured. Older SQL databases can be integrated into the system. Semi-structured and unstructured data can be loaded into a data lake.
2. **Ingestion and pre-processing**: Azure Synapse Analytics processing pipelines and extract, transform, load (ETL) processing can connect to data stored in Azure or third-party sources via built-in connectors. Azure Synapse Analytics supports multiple analysis methodologies that use SQL, Spark, Azure Data Explorer, and Power BI. You can also use existing Azure Data Factory orchestration for the data pipelines.
3. **Processing:** Azure Machine Learning is used to develop and manage the machine learning models.
    - **Initial processing**: During this stage, raw data is processed to create a curated dataset that will train a machine learning model. Typical operations include data type formatting, imputation of missing values, feature engineering, feature selection, and dimensionality reduction.
    - **Training**: During the training stage, Azure Machine Learning uses the processed dataset to train the credit risk model and select the best model.
    - **- Model training**: You can use a range of machine learning models, including classical machine learning and deep learning models. You can use hyperparameter tuning to optimize model performance.
    - **- Model evaluation**: Azure Machine Learning assesses the performance of each trained model so you can select the best one for deployment.

SURAJIT SHOME
AZURE DATA ENGINEER
Phone: +31 647988248

    - **- Model registration**: You register the model that performs best in Azure Machine Learning. This step makes the model available for deployment.
4. **Responsible AI:** Responsible AI is an approach to developing, assessing, and deploying AI systems in a safe, trustworthy, and ethical way. Because this model infers an approval or denial decision for a loan request, you need to implement the principles of Responsible AI.
    - ***Fairness metrics*** assess the effect of unfair behavior and enable mitigation strategies. Sensitive features and attributes are identified in the dataset and in cohorts (subsets) of the data. For more information.
    - ***Interpretability*** is a measure of how well you can understand the behavior of a machine learning model. This component of Responsible AI generates human-understandable descriptions of the model's predictions. For more information.
5. **Real-time machine learning deployment:** You need to use real-time model inference when the request needs to be reviewed immediately for approval.
    - Managed machine learning online endpoint. For real-time scoring, you need to choose an appropriate compute target.
    - Online requests for loans use real-time scoring based on input from the applicant form or loan application.
    - The decision and the input used for model scoring are stored in persistent storage and can be retrieved for future reference.
6. **Batch machine learning deployment**: For offline loan processing, the model is scheduled to be triggered at regular intervals.
    - Managed batch endpoint. Batch inference is scheduled and the result dataset is created. Decisions are based on the creditworthiness of the applicant.
    - The result set of scoring from batch processing is persisted in the database or Azure Synapse Analytics data warehouse.
7. **Interface to data about applicant activity**: The details input by the applicant, the internal credit profile, and the model's decision are all staged and stored in appropriate data services. These details are used in the decision engine for future scoring.
    - Storage: All details of credit processing are retained in persistent storage.
    - User interface: The approval or denial decision is presented to the applicant.
8. **Reporting:** Real-time insights about the number of applications processed and approve or deny outcomes are continuously presented to managers and leadership. Examples of reporting include near real-time reports of amounts approved, the loan portfolio created, and model performance.

Components to be used

Azure Blob Storage provides scalable object storage for unstructured data. It's optimized for storing files like binary files, activity logs, and files that don't adhere to a specific format.

Azure Data Lake Storage is the storage foundation for creating cost-effective data lakes on Azure. It provides blob storage with a hierarchical folder structure and enhanced performance,

management, and security. It services multiple petabytes of information while sustaining hundreds of gigabits of throughput.

Azure Synapse Analytics is an analytics service that brings together the best of SQL and Spark technologies and a unified user experience for Azure Synapse Data Explorer and pipelines. It integrates with Power BI, Azure Cosmos DB, and Azure Machine Learning. The service supports both dedicated and serverless resource models and the ability to switch between those models.

Azure SQL Database is an always up-to-date, fully managed relational database that's built for the cloud.

Azure Machine Learning is a cloud service for managing machine learning project lifecycles. It provides an integrated environment for data exploration, model building and management, and deployment and supports code-first and low-code/no-code approaches to machine learning.

Power BI is a visualization tool that provides easy integration with Azure resources.

Azure App Service enables you to build and host web apps, mobile back ends, and RESTful APIs without managing infrastructure. Supported languages include .NET, .NET Core, Java, Ruby, Node.js, PHP, and Python.

Azure Databricks to develop, deploy, and manage machine learning models and analytics workloads. The service provides a unified environment for model development.