Cheat Sheet: Data Engineering and Generative AI

Popular data engineering and AI tools

Tool	Usage	Link
DataRobot	Automated machine learning and AI	www.datarobot.com
Prefect	Data workflow management	www.prefect.io
Hazy	Synthetic data generation	https://hazy.com/
ChatGPT	AI language model	https://openai.com/chatgpt
StitchData	Data integration	www.stitchdata.com
Schemawriter	Data schema generation	schemawriter.ai
universaldata.io	Data Generation and Augmentation	https://www.universaldata.io/
DBdiagram.io	Draw ER diagram creation	https://dbdiagram.io/

Important prompts for generative AI for architecture design

Task	Prompt
Generate data architecture design for a hospital network	Create a detailed data architecture design for a hospital network.
Add data modeling components for patient data	Create additional data modeling components for patient demographics, medical history, diagnosis, treatment, and quality standards.
Design to implement robust access controls, data privacy, and audit mechanisms	Create a design to include implementing robust access controls, encryption, and auditing mechanisms to protect patient data from unauthorized access or breaches.
Detailed data architecture design for a retail company's customer relationship	Create a detailed data architecture design for a retail company's customer relationship management system.
Generate unified customer profile	Create a unified customer profile with attributes such as demographics, purchase history, browsing behavior, preferences, and interactions.

Important prompts for generative AI for database, data warehouse schema design

Task	Prompt
Generate data warehouse schema for a fashion retail store	Create a detailed data warehouse schema for a fashion retail store that should contain: i) Employee data ii) Sales data iii) Inventory data iv) Customer profiles v) Seller information

Important prompts for generative AI for data anonymization

Task	Prompt
Anonymize names	Replace the entries under the 'Name' attribute of a data set with pseudonyms like "User_i" using Python.
Redact email addresses	Write a Python code to redact the entries under the attribute 'Email' in a data frame so that only the usernames and service provider's first and last characters are visible. Rest all characters are replaced with the character '*'.
Generalize ages to decades	Write a Python code to generalize the entries under the attribute 'Age' of a data frame such that the exact number is converted into a generic range
Add noise to contact numbers	Write a Python code to add random noise of 5-digit length to a numerical attribute 'Contact Number' in a data frame with all ten-digit length values.

Important prompts for generative AI for infrastructure setup

Task	Prompt
Generate the e-commerce platform's data infrastructure	Write Python code that performs the following tasks: Create the steps for the e-commerce platform to enhance its data infrastructure to handle the increase in traffic. Suggest the improvements in 1. Scalable storage 2. Better processing capabilities 3. Real-time analytics
Generate data lake infrastructure for a healthcare platform	Write Python code that performs the following tasks: Create the steps for a healthcare company to set up a data lake infrastructure that is capable of the following: 1. Big data management 2. Data ingestion from various sources 3. Data transformation 4. Data security and compliance with regulatory guidelines
Generate infrastructure to detect real-time fraudulent transactions for a financial firm	Write Python code that performs the following tasks: Create the steps for a financial firm to set up its infrastructure if they want to detect fraudulent transactions in real time. Suggest specifics in terms of: 1. Computing machinery 2. Feature engineering pipeline 3. Predictive modeling pipeline 4. Model deployment and monitoring

Author(s)

Skills Network