Yanna Shen

206-290-3069 yanna.cshen@gmail.com www.linkedin.com/in/chengyang-shen https://yannacs.github.io/

SUMMARY

Data Engineer with strong foundation in building scalable data pipelines, ETL workflows, and cloud-based data solutions. Experienced in Python development, database design, API engineering, and big data technologies including Apache Spark and streaming architectures. Proven track record of designing and deploying production-ready data infrastructure, implementing real-time data processing systems, and optimizing data workflows for performance and reliability.

EDUCATION

Northeastern University Seattle, WA
Master of Science in Data Analytics Engineering (GPA: 3.96 / 4.00) *Sept. 2023 - May 2025*
Xiamen University Malaysia
Bachelor of Engineering in Computer Science and Technology (Honors) *Sept. 2017 - Aug. 2021*

SKILLS

Programming Languages: Python, SQL (MySQL, PostgreSQL, NoSQL, MongoDB), Shell Scripting, JavaScript, Java
Data Engineering: ETL/ELT Pipelines, Data Modeling, Data Warehousing, Data Lake Architecture, Data Integration, Data Quality, Stream Processing, Batch Processing, Data Governance, AWS (S3, EC2, Lambda, RDS, Glue, Redshift, Kinesis), Docker, CI/CD, Git, Linux/Unix, Infrastructure as Code, Apache Spark (PySpark), Kafka, Hadoop Ecosystem, Distributed Computing, Real-time Data Processing, Event-Driven Architecture
Database Technologies: Relational Databases (MySQL, PostgreSQL), NoSQL (MongoDB, Redis), Database Design, Query Optimization, Indexing, Database Normalization, ORM (SQLAlchemy, Django ORM)
API Development: RESTful APIs, FastAPI, Django REST Framework, Flask, GraphQL, Authentication (JWT, OAuth), API Gateway, Microservices
Data Tools & Libraries: Pandas, NumPy, Apache Airflow, dbt, Great Expectations, PySpark, JDBC, SQLAlchemy
Visualization & Analytics: Tableau, Streamlit, Matplotlib, Business Intelligence, Data Analysis, Statistical Analysis
Machine Learning: Scikit-learn, TensorFlow, Feature Engineering, Model Deployment, MLOps Fundamentals
Certifications: Tableau Certified Data Analyst, AI Literacy, Data Analytics on AWS, Applied AI

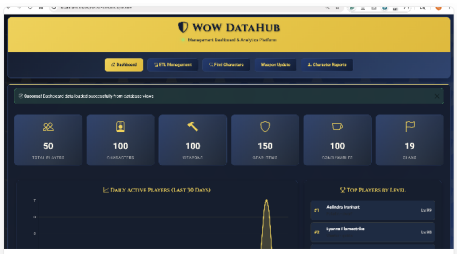
EXPERIENCE

Data Engineering Intern | **Uplift Northwest, Seattle** *(ETL, Data Pipeline, Cloud)* June - Sept. 2024

- Architected and implemented scalable ETL pipelines integrating multiple disparate data sources, processing over 100K records daily with automated error handling and data validation
- Designed normalized relational database schemas and optimized data models, reducing query execution time by 40% through strategic indexing and query optimization
- Built cloud-ready data warehouse infrastructure on AWS, implementing data partitioning strategies and incremental loading patterns for improved performance
- Developed automated data quality monitoring framework with real-time alerting, reducing data issues by 60% and improving overall data reliability
- Created interactive Tableau dashboards integrated with data pipelines, enabling real-time business intelligence and stakeholder decision-making

Data Engineer | **Zhejiang Earthview Image Inc., China** *(Big Data, ETL, Cloud Infrastructure)* Sept. 2020 - May 2022

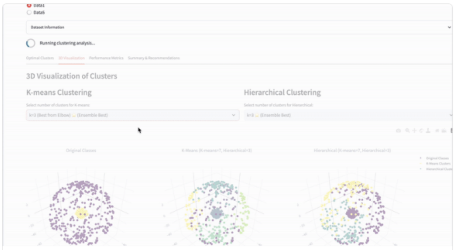
- Engineered a high-throughput data processing pipeline for satellite imagery using Apache Spark and Hadoop, handling terabytes of geospatial data and enabling faster image analysis for downstream applications
- Implemented robust ETL workflows for remote-sensing data ingestion, transformation, and storage in an AWS S3 data lake, improving data availability for analytics teams
- Designed and deployed automated data quality validation system, ensuring 99.5% data accuracy across multiple data sources
- Built real-time analytics infrastructure supporting A/B testing framework, processing millions of user events daily for platform optimization
- Optimized data storage and retrieval patterns, reducing storage costs by 30% while improving query performance through efficient data partitioning
- Developed computer vision data pipelines integrating ML models with production data systems, achieving 87% classification accuracy at scale



WoW DataHub ☞ Game Management Dashboard & Analytics Platform

A comprehensive database solution for MMORPG character data management, featuring:

- Normalized relational database with optimized schema design
- Robust ETL pipeline for automated data ingestion and transformation
- Secure JAVA-based data access layer with SQL injection prevention
- Complex analytics engine for multi-dimensional game statistics
- Interactive web dashboard for stakeholder insights

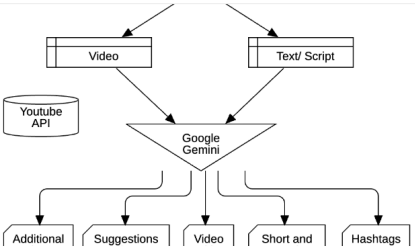


Clustering Analysis and 3D Visualization Dashboard

A dashboard for analyzing datasets by using K-means and Hierarchical clustering algorithms. This tool helps determine the **optimal number of clusters** and provides detailed **3d visualizations** and comparisons of clustering results.

🎯Click here and Try it out

(plz get it back up and wait it waking up for a moment)



Co-Creator ☞ AI-Powered Video Creation Assistant

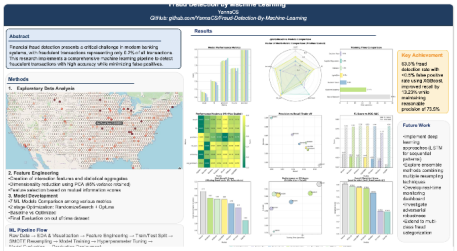
An intelligent video production companion that streamlines the entire content creation workflow using Gemini AI technology. This comprehensive tool assists creators at every stage of video production, from initial concept to final optimization.

🎯Click here and Try it out

(plz get it back up and wait it waking up for a moment)

Type 2 Diabetes Risk Prediction Model

This project develops machine learning models to predict diabetes status and identify key risk factors using survey data from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) 2022 dataset, containing 445,132 participants across 328 variables.



Fraud Detection by Machine Learning

A comprehensive machine learning pipeline for detecting fraudulent transactions with 99% accuracy. This project tackles extreme class imbalance using advanced techniques including SMOTE, ensemble methods, and automated hyperparameter optimization to protect financial systems in real-time.