

## Basics

- Senior Data Engineer with 6+ years of experience specializing in distributed big data systems and cloud-native architectures, well-positioned to support high-scale Automated Driving (AD) R&D environments.
- Proficient in AWS and Azure ecosystems, utilizing Spark, Python, and Kubernetes to design robust backend applications and CI/CD workflows for enterprise-grade data processing.
- Proven ability to lead technical initiatives and define big data strategies, bridging the gap between complex business requirements and efficient, scalable infrastructure solutions.

## Education

- Sep 2015–Jul 2019 **Bachelor, Computer Science and Technology, Suzhou University of Science and Technology**
- Awarded Third Prize in the 4th National University Cloud Computing Application Innovation Competition (2018).
  - Project: "Implementation of Large-scale Distributed Functional Dependency Mining Algorithm Based on Spark".

## Work Experience

- Dec 2022–Present **Senior Data Engineer, Cognizant Technology Solutions Shanghai Co. Ltd.,** <https://www.cognizant.com>
- Led the migration of a 10TB+ legacy data warehouse to AWS, utilizing AWS Glue and DMS to modernize infrastructure, resulting in a 40% reduction in query latency and 25% lower maintenance costs.
  - Architected a unified Customer Data Platform (CDP) integrating 8+ disparate systems (Salesforce, Databricks) using Spark and Python, processing billions of records daily to enable a 360-degree customer view.
  - Optimized core data processing workflows in Databricks by implementing advanced partitioning and caching strategies, reducing daily processing time from 6 hours to 2.5 hours (58% improvement).
  - Designed a Master Data Management (MDM) solution on Azure Data Lake to resolve critical retail data inconsistencies, reducing discrepancies by 95% and streamlining operations.
- Keywords:** AWS (Glue, DMS), Databricks & Spark, Python, Cloud Migration, Data Architecture
- Jun 2021–Nov 2022 **Data Engineer, Patsnap Information Technology (Suzhou) Co. Ltd.,** <https://www.patsnap.com>
- Architected a scalable data lake solution on AWS S3 and Athena for 1.8 billion global patent records, orchestrating Spark jobs via DolphinScheduler to support TB-scale financial analysis.
  - Developed a custom "Spark Diff" engine to synchronize 500M+ daily records to TiDB, achieving minute-level processing for complex nested structures and ensuring efficient incremental updates.
  - Built a real-time wide table system using Flink, Kafka, and TiCDC, patching open-source components to enable sub-second data ingestion for enterprise business insights.
  - Designed "Guardian," a continuous delivery tool integrated with GitLab CI/CD, automating Docker deployments and reducing delivery costs for 10+ localized banking projects.
- Keywords:** AWS (S3, Athena, EMR), Apache Spark & Flink, Kafka, CI/CD (GitLab), TiDB
- Jun 2019–Jun 2021 **Data Engineer, Intsig Information Co., Ltd.,** <https://www.intsig.com>
- Engineered a high-throughput real-time data pipeline using Redis and Kafka, processing over 100 billion data points to ensure sub-second freshness for a platform with 230M+ entities.
  - Architected a Python-based ETL optimization library that standardized incremental update logic across 1000+ dimensions, reducing boilerplate code by 90%.
  - Led the DevOps transformation by migrating from SVN to GitLab and implementing CI/CD pipelines, significantly improving team development efficiency and deployment speed.
- Keywords:** Real-time Streaming, Kafka & Redis, Python ETL, DevOps, Data Modeling

## Skills

Cloud & Infrastructure Expert **Keywords:** AWS (EMR, Glue, Redshift), Azure (Synapse, Data Lake), Kubernetes & Docker, Terraform (IaC), CI/CD (GitLab, Jenkins)

Big Data & Streaming	Expert	<b>Keywords:</b> Apache Spark & Flink, Kafka Event Streaming, Hadoop Ecosystem (Hive), Databricks Platform
Data Architecture & Engineering	Advanced	<b>Keywords:</b> Data Warehouse & Lakehouse, ETL/ELT Pipeline Design, Serverless Data Processing, Airflow Orchestration
Programming & Languages	Advanced	<b>Keywords:</b> Python & SQL, Java & Scala, Shell Scripting
Database Systems	Advanced Server	<b>Keywords:</b> PostgreSQL & MySQL, NoSQL (MongoDB, Redis), TiDB (NewSQL), SQL

## Certificates

Jan 2024	<b>AWS, AWS Certified Data Analytics –Specialty</b>
Sep 2024	<b>AWS, AWS Certified Developer - Associate</b>
Jun 2024	<b>Microsoft, Microsoft Certified Azure Data Engineer Associate</b>
Mar 2023	<b>Databricks, Databricks Certified Data Engineer Associate</b>
Mar 2022	<b>PingCAP, PingCAP Certified TiDB Professional</b>

## Projects

Dec 2022– Present	<b>Cloud-native data lakehouse unifying customer data across the enterprise., Bayer Customer Data Platform (CDP)</b> <ul style="list-style-type: none"> <li>○ Architected a serverless ETL architecture using AWS Glue and Lambda, optimizing compute costs by 40% while maintaining strict SLAs for data freshness.</li> <li>○ Established a robust Infrastructure as Code (IaC) foundation using Terraform and GitHub Actions, reducing deployment time by 60% and ensuring 99.9% infrastructure uptime.</li> <li>○ Designed an event-driven orchestration workflow with AWS Step Functions and EventBridge to manage complex dependencies across MySQL, SQL Server, and PostgreSQL.</li> </ul> <b>Keywords:</b> AWS (Glue, Lambda, Redshift), Terraform, GitHub Actions, Data Lakehouse
Nov 2022– May 2022	<b>Strategic migration of a 10TB+ legacy warehouse to Amazon Redshift., Enterprise Data Warehouse Migration to AWS</b> <ul style="list-style-type: none"> <li>○ Developed a custom Python migration engine () using AST parsing () to achieve 99% automated DDL translation accuracy for over 2,000 tables.</li> <li>○ Engineered a high-throughput stream processing pipeline using UNIX named pipes and , accelerating 10TB data transfer by 300% compared to traditional staging.</li> <li>○ Designed performant Redshift schemas with optimal Distribution and Sort Keys, resulting in a 5x improvement in complex analytical query performance.</li> </ul> <b>Keywords:</b> Amazon Redshift, Python, Database Migration, Performance Tuning
May 2025– Dec 2025	<b>Centralized retail data governance system on Microsoft Power Platform., Retail Master Data Management Platform</b> <ul style="list-style-type: none"> <li>○ Designed a billion-scale storage layer using Azure Cosmos DB, ensuring sub-second response times for current state queries while maintaining full historical versioning.</li> <li>○ Implemented a cross-cloud data synchronization bridge between Azure Dataverse and AWS analytical warehouses using Synapse Link and custom adapters.</li> </ul> <b>Keywords:</b> Azure Cosmos DB, Power Platform, Data Governance, Cross-Cloud Integration