

Jan Krol

Enterprise Data Expert

Ein Bild, das Text, blau, Schild enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Schild, blau enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Schild enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Schild enthält.

Automatisch generierte BeschreibungEin Bild, das Text, Schild enthält.

Automatisch generierte Beschreibung

|  |  |
| --- | --- |
| Personal Information |  |
|  | **Big Data Specialist**    Focus: Big Data, Cloud Architecture, Data Management Platforms |
|  | **Professional Experience:**   * Deutsche Bahn * Jungheinrich * Lufthansa Technik * Fielmann * BMW Group |
|  | **Language Skills:**  German –Native Speaker English – Business Fluent  Polish – Business Fluent |

|  |  |
| --- | --- |
| Skills | * Big Data Platform Specialist with focus on Amazon Web Services & Microsoft Azure * ETL Processes/Pipelines & Data Engineering * Architecture of Data Management Plaftorm in Enterprises * Build up of Data Lakes & Data Lakehouses * Application migrations using cloud services * Consulting & implementation of automation concepts especially DevOps * Integration of Active Directory, security concepts and compliance requirements * Monitoring and Logging * Confident in Python, SQL, TypeScript, Golang |

|  |  |
| --- | --- |
| Methodological and Technical Focus | * Big Data Cloud Architecture (AWS & Microsoft Azure) * Data Engineering (Databricks, Synapse Analytics, Fabric, Apache Spark, AWS Glue, Athena, Redshift & EMR) * Infrastructure as Code (Terraform, Pulumi, AWS CDK, ARM) |

|  |  |
| --- | --- |
| Certifications | * Azure Solutions Architect Expert:   + AZ-300: Microsoft Azure Architect Technologies   + AZ-301: Microsoft Azure Architect Design * Databricks Certified Associate Developer   + for Apache Spark 3.0 * HashiCorp Certified: Terraform Associate * AWS Certified Machine Learning – Specialty * AWS Certified Solutions Architect – Associate * AWS Certified Cloud Practitioner * AWS Technical Professional * AWS Business Professional |
| Projects  **03/2023**  **| @ Intralogistics**  **05/2024**  **01/2022**  **| @ Logistics**  **02/2023**  **09/2021**  **| @ E-Commerce**  **01/2022**  **04/2020**  **| @ E-Commerce**  **09/2021**  **02/2019**  **| @ Transport & Logistics**  **04/2020**  **09/2018**  **| @ Transport & Logistics**  **02/2019**  **Before 2018** | * **Development of a Comprehensive Data Strategy and Governance Framework for an Advanced Data Management Platform on Databricks** * ***Project Summary****:*   *In this pivotal project, I spearheaded the development of a robust data strategy and governance framework, aiming to streamline and enhance the organization's data handling capabilities. The core of this project involved the construction of a sophisticated data management platform on Databricks, complemented by the design and implementation of an efficient data hub ingestion platform.*   * + Led the design and establishment of an organization-wide data strategy, aligning it with business goals and technological advancements   + Developed a comprehensive data governance framework to ensure data accuracy, privacy, and compliance with industry standards   + Oversaw the deployment and customization of the data management platform on Databricks, enhancing data processing, analysis, and reporting capabilities with Power BI   + Engineered a robust data hub with advanced ingestion pipelines based on AWS EventBridge, optimizing data flow from diverse sources to centralized storage systems (Data Lake House on Azure)   + Collaborated with cross-functional teams to integrate the data management platform with existing IT infrastructure and business processes   + Conducted training sessions and workshops for staff, fostering a data-driven culture and enhancing data literacy across the organization   + Services:     - **Azure Databricks**     - **Databricks Data Catalog**     - **AWS EventBridge**     - **Kinesis**     - **Event Hub**     - **Structured Streaming (Apache Spark)** * **Advanced Integration and Analysis of Logistics Data Streams with PySpark Structured Streaming and Data Mesh Implementation** * ***Project Summary:***   *This project focused on the sophisticated integration of logistics data streams with Event Hub and Kafka using PySpark Structured Streaming. Our approach revolutionized the way logistics data is captured, processed, and analyzed in real-time. Leveraging advanced technologies such as GraphFrame, Azure Synapse Analytics, Apache Spark, and PowerBI, we established a robust system that not only ensures high data quality and smooth transmission but also aligns with IT Governance principles and utilizes the Data Mesh paradigm.*   * + Integration of a logistics data stream with Kafka through PySpark Structured Streaming   + Design and implementation of a pipeline for capturing, processing, and forwarding the data stream   + Utilizing PySpark Structured Streaming for efficient data processing   + Services:     - **Azure Synapse Analytics**     - **Purview Data Catalog**     - **Apache Spark**     - **Event Hub**     - **Structured Streaming**     - **GraphFrame**     - **Azure Storage v2**     - **PowerBI**   + Configuration and initialization of the PySpark Streaming job   + Defining the necessary data structures for the data stream   + Conducting comprehensive testing and monitoring mechanisms   + Ensuring smooth data transmission and high data quality   + Robust and efficient integration of the logistics data stream with Event Hubs   + Real-time utilization of logistics data for analysis and further processing * **Enhanced Data Processing and Integration Systems for E-commerce with Serverless and Distributed Data Mesh Architectures** * ***Project Summary:***   *In this project, my primary role was to guide and support various internal e-commerce product teams in developing, implementing, and maintaining high-performance data processing and integration systems. The focus was on migrating existing data services and pipelines to a new, enhanced architecture, emphasizing the development of an event-based system using serverless technologies and Big Data frameworks.*   * + Support and guidance in migrating existing data services, pipelines, and assets to a new and enhanced architecture   + Development of an event-based system   + Utilization of Lambda functions and PySpark   + Integration with Kafka   + Design and architecture planning   + Implementation of Lambda functions and PySpark jobs   + Configuration and connection with Kafka   + Serverless architecture for scalability and availability   + Real-time processing and analysis of event data   + PySpark transformations, filtering, and aggregations   + Efficient and reliable connection with Kafka   + Configuration, security settings, and integration with other components   + Extensive testing and monitoring mechanisms   + High-performance and scalable event system   + Extraction of valuable insights from event data   + Data-driven decision-making   + Services:     - **AWS Glue**     - **Apache Spark**     - **Data Catalog**     - **S3**     - **Athena**     - **Redshift**     - **Lambda**     - **ECS**     - **Step Functions**   + Implementing distributed data mesh architectures to enable different product teams to work efficiently with data   + Data processing with Big Data frameworks and database technologies   + Development of serverless/elastic cloud architectures (AWS)   + Deploying the architecture according to DevOps best practices and Infrastructure-as-Code (AWS CDK & Terraform). * **Migration and Enhancement of E-commerce Analytics to AWS Data Lakehouse Architecture** * ***Project Summary:***   *This project involved the strategic development and migration of existing analytics data pipelines into a sophisticated Data Lakehouse architecture, leveraging a range of AWS services. A key aspect was the enhancement of our Big Data Lake environment and ensuring stringent data quality and compliance standards, particularly with GDPR.*   * + Further development of the Big Data Lake environment in AWS   + Implementation and conception of a Data Lakehouse   + Exploratory analysis as well as algorithm development through data provisioning and preparation (AWS Glue, Spark, Lambda)   + Data ingestion   + Development of data pipelines, ETL jobs to provide ready-to-consume data sources (AWS Glue, AWS Redshift, Spark, PySpark)   + Regression testing and quality checking in the data pipelines and data lake   + High Performance Streming Data Processing (Kinesis / Kafka) + Lambda   + Orchestration and connectivity of data sources   + Implementing automated deployments using DevOps best practices (AWS Codebuild + Codepipeline, GitHub Actions)   + Building the infrastructure using IaC (AWS CDK)   + Data quality and compliance (incl. cost monitoring)   + Services:     - **AWS Glue**     - **Kinesis**     - **Kafka**     - **Apache Spark**     - **Data Catalog**     - **S3**     - **Athena**     - **Redshift**     - **Lambda**     - **ECS**     - **Step Functions** * **Development of architectures and implementation in a big data environment for group-wide, standardised platform services based on Azure** * ***Project Summary:***   *This project involved developing and implementing a standardized big data architecture for group-wide platform services in the Transport & Logistics sector, utilizing various Azure services. My role was crucial in ensuring the integration of data transparency, adherence to Data compliance, and the implementation of agile methodologies.*   * + Development and automation of solutions in sub-projects on Azure and presentation/ discussion of these   + Building infrastructure with Infrastructure as Code (Terraform) and Ansible   + Scrum, JIRA, GIT, Docker   + Implementation of real-time data streaming with Apache Kafka   + Monitoring   + Consulting for an Azure platform strategy about reference architectures   + Development of mechanisms and automatisms for the proactive elimination of vulnerabilities of the Azure and Kubernetes components based on standardised clusters (security by default)   + Conceptual development of the architectural and technological platform in the field of container orchestration based on Kubernetes, Continuous Integration & Continuous Deployment   + Creation of user and authorisation concepts considering the group specifications   + Operational management of the services offered   + Agile team membership   + Services:     - **Azure Data Catalogue (Purview),**     - **Azure Synapse Workspace Analytics,**     - **Azure Data Factory,**     - **Azure Databricks**     - **Terraform**     - **Gitlab Runner**     - **Azure DevOps** * **Pioneering AWS Infrastructure Consultation and Implementation for Global Process Operations in Transport & Logistics** * ***Project Summary:*** * *This project entailed providing expert consulting and hands-on implementation of AWS infrastructure to support a process operations team responsible for several international applications in the transport and logistics sector. My role was crucial in identifying and implementing optimizations, developing and maintaining system infrastructure, and providing comprehensive support and training to internal teams.*   + Provision and operation of servers, operating system environments and database systems in AWS   + Recognition of optimization potential in both commercial and technical terms   + Development and presentation of optimized processes   + Implementation of optimizations   + Acting on one's own responsibility and representing the developed results within the team and towards the project managers/clients   + Administration and maintenance of the provided systems   + Development of maintenance and monitoring concepts for these systems   + Support and advice for development projects on the use, configuration and optimisation of the systems provided   + Consulting on architectures and operational concepts using the AWS Cloud infrastructures   + When using new AWS services: Training of internal employees for the changed working methods * **Implementation of a machine learning based analysis tool for automatic detection of noise in vehicle test data**   + ETL process design (**AWS Glue** | **AWS Redshift**)   + Provision of AWS infrastructure (AWS Glue + Redshift + EMR)   + Design and construction of the anomaly logic (AWS Sagemaker, Python)   + Realization of pre-processing of the test data   + Implementation of automatic anomaly detection to detect interference frequencies in scikit-learn (AWS Sagemaker)   + Visualization of the analysis results * **Application migration for a business unit (Transport & Logistics) incl. development of AD**   + Provision of AWS infrastructure: databases (SQL) & EC2 machines, AWS Lambda Services   + Deployment using Terraform   + Planning and execution of application migration   + Rolling out authorizations   + Infrastructure provisioning with AWS Cloudformation * **Implementation and technical project support for a web application to manage the certification process in consideration of compliance requirements for a well-known automobile manufacturer**   + Infrastructure deployment on AWS (MySQL Server)   + Front end development in React.js   + Back End realization with Java EE | glassfish   + Refactoring and code optimization   + Agile operation through Scrum with Jira   + CI/CD with Jenkins   + Documentation in Confluence * **Development and consulting of a marketplace platform based on Microsoft Azure Services**   + Integration Payment Provider   + Planning and architecture of Microsoft Azure Services   + Implementation/consulting for backend and frontend applications   + Creation of User Management (Azure AD)   + Implementation of an upload tool for very large files   + Security Engineering * **Implementation of a process automation based on a vulnerability assessment**   + Design and planning of the most cost-efficient infrastructure components   + Implementation of Python logic in AWS Lambda   + Provision of the infrastructure with AWS Cloudformation via YAML templates   + Code Optimization   + Sending notification emails via AWS SNS |
|  |  |
|  |  |
|  |  |