Raj Kumar Rai – Bergfreunde Technical Interview Preparation (v4: Data Engineering + AI Strategy)

Date: 29 October
Interviewer: Christopher Barth (Head of Data)
Focus: Enterprise Data Strategy · Spark & Declarative Pipelines · Databricks · AI Integration

1. Elevator Pitch (Balanced & Strategic)

I'm Raj Kumar Rai — currently Lead Data Engineer at Freeletics in Munich. With over 19 years of experience, I specialize in designing scalable data platforms and AI-enabling architectures.

At Freeletics, I led the design of an enterprise Databricks Lakehouse on AWS that handles both batch and streaming data at scale using Spark, Delta Live Tables, and declarative pipelines. I implemented Unity Catalog for governance, MLflow for MLOps, and GitHub Actions with Terraform for CI/CD automation.

My approach combines data engineering best practices — strong data models, pipeline reliability, and performance tuning — with AI strategy, integrating Mosaic AI and AgentBricks to create intelligent agents that leverage governed Delta data.

I see Bergfreunde at an exciting stage where a unified Lakehouse can connect engineering, analytics, and AI. My strength lies in building the foundations — reliable data pipelines, scalable architectures, and intelligent extensions — that make that vision possible.

2. Core Technical & Strategic Questions

Q1: Describe your Lakehouse architecture.

I architected a multi-layer Databricks Lakehouse using the Medallion design (Bronze–Silver–Gold) to separate ingestion, cleansing, and consumption layers. Bronze ingests raw data from APIs, Kafka, and Kinesis using Autoloader. Silver standardizes and enriches data using PySpark and Delta Live Tables for declarative transformations and quality enforcement. Gold provides aggregated business datasets supporting BI dashboards, ML models, and downstream systems. All pipelines are declarative, version-controlled, and governed by Unity Catalog — ensuring a trusted foundation for both analytics and AI.

Q2: How do you approach building robust and maintainable data pipelines?

I design for scalability, maintainability, and observability. Delta Live Tables allows declarative dependency management and data quality enforcement. Complex transformations are modularized with PySpark functions. Jobs run on autoscaling job clusters with Databricks metrics for observability, integrated into CI/CD via automated testing. This ensures resilient, production-grade pipelines that scale dynamically and recover gracefully from failures.

Q3: How do you align data engineering with AI initiatives?

AI success depends on strong data foundations. I design feature-ready pipelines that connect Silver/Gold datasets directly with MLflow and Feature Store, ensuring train-serve parity. Declarative data contracts validate schema changes automatically across data and ML layers, creating a smooth handoff from data to model to production. This unified design bridges engineering and data science, enabling faster and more reliable AI delivery.

Q4: How do you ensure cost efficiency and performance optimization?

I apply optimization across compute, storage, and code. Compute: autoscaling spot job clusters and workload segregation. Storage: tiering via Delta history retention to S3 Standard, IA, and Glacier. Code: Z-ordering, partition pruning, and caching to enhance Spark performance. These measures combined with cost dashboards achieved ~30% compute savings and transparent cost governance.

Q5: How do you collaborate across teams and scale data culture?

I establish reusable ingestion and transformation frameworks that let multiple teams contribute to the Lakehouse consistently. BI teams consume curated Gold tables; Data Science leverages the Feature Store for reproducibility. I promote DataOps through automated tests, lineage visibility, and code reviews, building a cross-functional culture grounded in trust and efficiency.

Q6: How do you integrate Generative AI in your architecture?

Generative AI extends the Lakehouse’s value. I integrate Mosaic AI and AgentBricks for AI orchestration and agent execution. Mosaic AI manages fine-tuning and inference, while AgentBricks enables agents to interact with Delta tables via SQL and Vector Search. Examples include metadata summarization agents, anomaly-report generators, and conversational data assistants. This transforms the Lakehouse into an intelligent, self-documenting, and insight-driven platform.

Q7: What trends excite you in Data & AI?

I’m particularly interested in the convergence of declarative data engineering and AI orchestration. Tools like DLT and LakeFlow simplify pipeline management, while Mosaic AI and AgentBricks bring intelligence to data operations. This synergy creates adaptive Lakehouses that self-optimize and democratize data and AI across the enterprise.

3. Generative AI & AgentBricks Case Studies (Integrated with Data Engineering)

1. Automated Metadata Enrichment
Developed an AgentBricks workflow connected to Unity Catalog that scanned Delta tables and generated structured metadata summaries via LLMs. Result: 80% reduction in manual curation effort.

2. Data Quality Summarizer
Built a Mosaic AI agent parsing Spark job logs and summarizing pipeline health, schema drifts, and anomalies. Shared daily insights with engineering teams.

3. Conversational Data Access
Integrated AgentBricks with Vector Search to let users query sales and marketing performance data in natural language, powered by Gold Delta tables.

4. Closing Vision

My strength lies in uniting data engineering discipline and AI innovation. I believe Bergfreunde can evolve its Lakehouse into a strategic data and AI platform — reliable, declarative, cost-efficient, and intelligence-ready.

With deep expertise in Spark optimization, declarative pipelines, and Databricks ecosystem integration, I can help scale Bergfreunde’s data foundation into a governed, AI-powered enterprise platform.