ID n° 13353

Data Reply GmbH - DE

PROFILE

Consultant.
As a seasoned Big Data Consultant with a Master's degree in Computer Science from the University of Bonn, where I worked on Automated selection of a RDF (Knowledge Graph) partitioner on the basis of RDF dataset statistics. I specialize in developing and optimizing data-driven solutions that enhance operational efficiency across diverse industries. My technical proficiency spans a broad array of tools and platforms, including Apache Spark, AWS, Azure, and Databricks, with certifications in AWS Solutions Architecture and Machine Learning. My recent role as a Data Engineer for a leading drink manufacturer involved innovative on-edge and cloud solutions to streamline data processes, showcasing my ability to integrate complex systems seamlessly. Throughout my career, I have successfully migrated legacy systems, reduced execution times, and improved data accessibility and analysis, significantly impacting business intelligence outcomes. I excel in deploying scalable data architectures, ensuring data integrity, and delivering actionable insights through customized dashboards and reports. Fluent in English and proficient in German, I effectively collaborate with global teams to drive strategic data-centric transformations. My goal is to continue leveraging my analytical skills and technical expertise to help organizations harness the power of big data for informed decision-making.

KEY SKILLS

Technical

- Apache Spark, Azure Databricks, Dataiku, Superset; Python, Scala, SQL; AWS (S3, EMR, Athena, Glue, EC2, Lambda, SageMaker); Azure (Resource Group, Storage Account, Event hub, PostgreSQL, Container Apps, Alerts, Key-Vault, DevOps); Git, Jira, Confluence, Docker; MySQL, SPARQL, Redshift, Snowflake; Terraform, Terragrunt

LANGUAGES

- **English**: Advanced (Speaking Proficiency); Advanced (Writing Proficiency)
- **German**: Intermediate (Speaking Proficiency); Intermediate (Writing Proficiency)
- **Hindi**: Native (Speaking Proficiency); Native (Writing Proficiency)

EDUCATION/TRAINING

- **04/2019**: *Master's - Computer Science (Big Data) - University of Bonn*

CERTIFICATION

- **2024**: Professional Scrum Master I - Scrum.org
- **2021**: Databricks Certified Associate Developer for Apache Spark 3.0 - Databricks
- **2020**: AWS Certified Machine Learning - Specialty - AWS
- **2019**: AWS Certified Solutions Architect - Associate - AWS

WORK EXPERIENCE

05/2024 -
Drink Manufacturer
Data Engineer

Description: Client needed a solution in factory to recognize whether empty bottles in received crates belong to their brand or to foreign brand.
Accomplishments:
- On Edge:
    - A camera is installed on edge to take crates images (top view) and further images are sent to model server for inference (detect no of brand bottles, no of foreign bottles, etc.).
    - RFID (with 2 antennas) device is installed on edge to read RFID tags on crates to detect crate epc_code etc.
    - Further, an app is deployed to join model inference with RFID output on timestamps to recognize how many brand bottles and foreign bottles a crate (epc_code) contains.
- On Cloud:
    - Development of data pipelines to ingest data coming from RFID devices, model output, etc. Pipelines includes:
        - Event Hubs (as queues) to ingest incoming data
        - Streaming apps to receive data from event hubs and load it to Postgres tables
        - Azure PostgreSQL to store data
        - Azure Container Apps to host streaming apps
        - Alerts to user group, if incoming data contains error or streaming apps stops working
    - Azure DevOps:
        - Azure Repos: For hosting Git repos
        - Azure Pipelines: To Build and deploy with CI/CD
        - Azure Boards: To follow agile methodology
- Terraform (Infrastructure as Code) and Terragrunt:
    - Automated deployment of Azure services using Azure Pipelines
    - Separate environments for development and production
- Power BI:
    - Power BI Desktop to create reports/dashboards on top of Postgres tables/views and publish them to Power BI Service
    - Power BI Service to share reports/dashboards with client

Skills: Python, Azure (Resource Group, Storage Account, Event hub, PostgreSQL, Container Apps, Alerts, Key-Vault, DevOps), Terraform, Power BI, Git

11/2022 - 04/2024
Vehicle Manufacturer
Superset Developer

Description: Client needed a data exploration and visualization application (Superset) where different business entities can see and analyze information related to ECUs (electric control unit) installed inside a vehicle instance. Also, dashboards needed to be optimized as they were failing due to the increasing amount of data and complex joins. Furthermore, the client required a data recovery strategy, in case of infrastructure failure and to make the application more robust.
Accomplishments:
- Ingestion of data from various sources to Superset for Creation of Superset dashboards.
- Deployment of dashboards through deployment scripts using VS code, to ensure dashboards recovery and rollbacks in case of any failures/errors.
- Dynamic querying of dashboards using Jinja2.
- Optimizations of SQL queries for existing dashboards.

- Designing Superset frontend according to project UI guidelines using CSS
Skills: Superset, Python, SQL, Jinja2, CSS

03/2021 - 10/2022
Clothing Company
Data Engineer

Description: In the past, the client used MS Access to prepare data workflows which were getting slower (because of the increasing amount of data) and error-prone (due to manual steps). A solution was needed to automate data preparation workflows. Furthermore, the client was interested in an automated workflow for inventory distribution. Also, an All-in-one platform was needed where different teams could see all business KPIs.
Accomplishments:
- Connection of various data sources to Dataiku for creating data workflows.
- Migration of data preparation workflows from Access to Dataiku by creating ETL pipelines on Dataiku.
- Optimization of workflows:
    o Reduced workflows execution time to 50-70 percent.
    o Reduced execution time to 90 percent for writing data to an external source.
- Creation of data pipelines in Dataiku for inventory distribution across different location clusters based on various KPI priorities, using PySpark and Python (Pandas).
- Creation and maintenance of ETL pipelines for developing All-In-One platform to see all business-related KPIs at a single interface.
- Development and maintenance of dataset validation tool to validate multiple business KPIs.
Skills: Dataiku, PySpark, Python, SQL, Power Query

11/2020 - 01/2021
Supermarket Chain
Data Engineer

Description: This supermarket chain was developing an application for online shopping. The data produced by the application was saved in SAP System. Amongst data storage, the goal was to have available different kinds of reports so further recommendations while online shopping can be performed to the clients.
The technical goal was to enable ETL jobs on SAP Data, to save intermediate results in Azure Blob Storage, and write the final reports on Snowflake.
Accomplishments:
- Reading and writing transformed data from/to Azure blob storage by running PySpark ETL jobs in Databricks notebooks.
- Data was further migrated to Snowflake for online shopping recommendations use case.
- Join optimizations to reduce running time for various bottlenecked Databricks notebooks, using different techniques such as bin packing algorithm, key salting technique, broadcasting and table indexing.
- POC for implementing incremental load of data in Databricks.
- Data quality checks were performed for various use case tables.
Skills: Databricks, Azure Blob Storage, Snowflake, PySpark, Python

08/2020 - 10/2020
Vehicle Manufacturer
Data Engineer

Description: The client's Analytics team wanted to analyze data coming from vehicles against different environmental factors. However, the data was not ready to be utilized by the analytics team as it was not clean and well structured. Data needed preprocessing and had to be made available for analysis.
Accomplishments:

- Developed various Glue Spark jobs pipelined together to filter/transform/preprocess data stored on S3 and created an Athena view on top of it for querying by the analytics team.
- Enable data access through Matlab for further analysis.
Skills: AWS (S3, Glue ETL Spark Job, Glue Crawlers, Athena), Spark, Python

01/2020 - 04/2020
Conglomerate Company
Data Engineer

Description: Client's analytics team needed traffic sensor data for making various business-related decisions. However, data received on S3 was near real-time data, wasn't in the best shape and structure to be queried directly. The project aim was to process, clean and aggregate traffic sensor data in near real-time and make it available for further analysis. Furthermore, the client wasn't sure which AWS services to be used for faster and efficient querying of data.
Accomplishments:
- Developed and executed automated PySpark jobs after every predefined time interval on EMR clusters to clean/process/transform/aggregate sensor data kept on S3. Output data gets stored on S3 again.
Skills: AWS (S3, EMR, SQS, SNS, Lambda, CloudWatch), Python, Spark, Jupyter notebook
- Output data from S3 was migrated to other AWS Services for accessing through Tableau so that the analytics team can perform analysis on it.
Skills: AWS (S3, Glue, Athena, Redshift, Redshift Spectrum), Tableau
- Performed a POC to compare the performance of queries executed through Tableau, when 1) Data is kept on Redshift Cluster 2) Data accessed through Redshift Spectrum and 3) Data accessed through Athena.
Skills: AWS (Redshift, Redshift Spectrum, Athena), Tableau

07/2019 - 12/2019
Sports brand
Data Engineer

Description: Due to the increasing amount of data, the client's previous solution (Alteryx) to host data pipelines wasn't performing well (slow and long execution times). The client was looking to migrate existing Alteryx workflows onto Big Data Platform such as Spark. Furthermore, the client was looking to develop new data pipelines for various business platforms so that various teams can utilize generated reports to make business decisions.
Accomplishments:
- Migration of Alteryx workflows to Spark/Scala scripts.
- Performance optimization of Spark/Scala jobs.
- CI/CD of processed data through Jenkins pipelines to Exasol for further analysis.
- Development of data pipelines for the various business platforms.
Skills: Spark, Scala, Jenkins, Hive, Exasol, Alteryx, SQL, AWS (EMR, S3)

06/2016 – 04/2016
Master Thesis
Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn (Germany)

Working on SANSA-Stack (big data processing engine for scalable processing of large-scale RDF data) on "Recommendation System for RDF Partitioners" using Spark and Scala

- Automated selection of a RDF (Resource Description Framework) Knowledge Graph partitioner on the basis of RDF dataset statistics. Using and comparing various algorithms supported by Spark's Machine Learning library (MLlib) such as Decision Tree, Random Forests, Gradient-Boosted Trees, etc. to recommend a partitioner.