ID n° 13353

Data Reply GmbH - DE

PROFILE

Consultant.
The Consultant is a Big Data Developer with over 4 years of academic and professional experience in working with big data technologies, especially Spark and AWS Cloud.

He is certified as Databricks Certified Associate Developer for Apache Spark 3.0, AWS Solutions Architect – Associate and AWS Machine Learning – Specialty.

During his M.Sc., he worked on a big data processing engine for scalable processing of large-scale RDF (Resource Description Framework) data and created a Recommendation System for RDF Partitioners using Spark and Scala.

Professionally, he worked with a sports brand for migrating the data preparation logic constructed through Alteryx workflows to Spark/Scala scripts.

Later, he worked for a large manufacturing company for cleaning/processing/transforming sensor data by writing PySpark jobs. Output data was kept on S3 and made availiable to Tableau server using various AWS services so that it can be analyzed by analytics team.

Further, he worked with a Vehicle manufacturer company for filtering/transforming data stored on S3 using Glue Spark jobs.

Recently, he worked for a retail company whose end goal was to develop an app for online shopping. His task was to prepare data provided by SAP system via writing various ETL jobs for online shopping recommendation use case.

Currently, he is working with a clothing company to create/maintain their data pipelines in Dataiku. His tasks also include migration of data workflows from Access to Dataiku and optimizations of Dataiku workflows.

KEY SKILLS

Technical

- **DBMS**: MySQL, MS SQL Server, SPARQL, HBase, Redshift
- **Development Tools**: IntelliJ IDEA, GitHub, Jira, Confluence, Bitbucket, Maven, Eclipse
- **Operating Systems**: Windows, MacOS, Linux
- **Platforms**: AWS, Apache Spark, Dataiku, Databricks, Azure, Snowflake
- **Programming Languages**: Python, Scala , JAVA, SQL

Soft

- Team player, fast learner, hard-working, adaptability

LANGUAGES

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

- **Hindi**: Native Speaker (Speaking Proficiency); Native Speaker (Writing Proficiency)

EDUCATION/TRAINING

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

CERTIFICATION

- **2020**: AWS Certified Machine Learning - Specialty - AWS
- **2020**: Databricks Certified Associate Developer for Apache Spark 3.0 - Databricks
- **2019**: AWS Certified Solutions Architect - Associate - AWS

WORK EXPERIENCE

03/2021 - 09/2022
Clothing Company Munich Germany
Data Engineering

Customer goal was to create/maintain data preparation workflows on Dataiku. Key task included:
- Creation of ETL pipelines on Dataiku:
o Migration of data preparation workflows from Access (slow and error prone) to Dataiku
- Optimization of workflows:
o Reduced workflows execution time to 50-70 percent
o External data source write optimization. Reduced data writing execution time to 90 percent.
- Creation of data pipelines in Dataiku for inventory distribution across different location clusters using PySpark and Python (Pandas)
- Developed and maintenance of dataset validation tool to validate multiple KPIs
- Creation and maintenance of ETL pipelines for developing All-In-One platform

Skills: Dataiku, PySpark, Python, SQL, Power Query

11/2020 - 01/2021
Supermarket Chain Munich Germany
Data Engineering

Customer goal was to develop an online shopping recommendation system. Key task included:
- Reading and writing transformed data from/to Azure blob storage by running PySpark ETL jobs in databricks notebooks.
- Data is further migrated to Snowflake for online shopping recommendataions use case
- Running time optimization for bottlenecked databricks notebooks
o Join optimizations using different techniques such as bin packing algorithm, key salting technique, broadcasting and table indexing.
o Reading big excel files efficiently and faster
- 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 Munich Germany
Data Engineering

Customers Analytics team wanted to analyze AdBlue consumption against different environmental factors. Data needed preprocessing and had to be made available for analysis. Key tasks included:
- Develop and execute several Glue Spark jobs pipelined together to filter/transform/preprocess data stored on S3 and create an Athena view on top of it for querying by the analytics team

- Enable data access through Matlab for further analysis

Skills: S3, Glue ETL Spark Job, Glue Crawlers, Athena, Spark, Python

01/2020 - 03/2020
Conglomerate Company Munich Germany
Data Engineering

The project aim was to process, clean and aggregate traffic sensor data and make it available to the internal analytics team for further analysis. Key tasks included:
- 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 anaytics 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 Munich Germany
Alteryx Migration

The customer was looking to migrate existing Alteryx workflows on to the new Big Data Platform as Spark jobs. Key tasks included:
- Migration of Alteryx workflows to Spark/Scala scripts
- Alteryx is a data preparation and data analytics tool. Alteryx workflows were written in Spark/Scala jobs
- Performance optimization of Spark/Scala jobs
- Data preparation for the Company’s trading platform
- Data preparation for the RDP (Rolling Demand Plan) platform of the company

Skills: Spark, Scala, Jenkins, Hive, Exasol, Alteryx, SQL, AWS (EMR, S3)

08/2013 - 09/2016
Multiple Gurgaon India
Patent data extraction

Worked on Big Data project using Hadoop framework

- To extract relevant patent data using MapReduce from a big data store

- To obtain total number of patents falling under one particular technological area Patent Landscape - Performed a few technical analysis to identify new ideas and helped clients to file patents. Other patent related activities

- Patent monetization & Infringement, patent invalidity, prior art search (PAS) and freedom-to-operate (FTO) searches.