dotmodus

Data Engineer

Scope:

To develop big data solutions using Google Cloud Platform. Design, build, implement, QA and deploy ETL transformations to enhance the Data Eco System and workflows. Responsible for data transformations - taking data from various apis, on premise transactional systems or alternative cloud providers and transforming this data so that it is in a presentable format for analysis, machine learning exploration and more traditional data visualisation methods.

- Design, build and maintain custom ETL processes
- Manage the Data Eco System – Design, implement, QA and deploy ETL transformations to enhance the Data Eco System and workflows
- Ensure data quality is maintained throughout all managed systems, perform data quality analysis and introduce monitors and alerts to maintain it
- Data Engineering Efficiency - Help ensure best practices and technologies are used in the department by constantly testing the data for corruption
- Build infrastructure to automate extremely high volumes of data delivery and creatively solve data volume and scaling challenges. Contribute to the design and architecture of innovative solutions to difficult problems
- Build data pipelines to pull together information from different source systems; integrating, consolidating and cleansing data; and structuring it for use in individual analytics applications
- Provide data in a ready-to-use form to data scientists who are looking to run queries and algorithms against the information for predictive analytics, machine learning and data mining purposes
- Work with business units and departments to deliver data aggregations to executives, business analysts and other end users for more basic types of analysis to aid in ongoing operations
- Apply various approaches to data architecture and applications to deal with structured and unstructured data sets
- Manage cloud workflow and production environment
- Manage databases and build out systems
- Design and produce model environments
- Develop machine learning capabilities
- Deliver models via the development of APIs
- Identify and build appropriate models to support client’s and user’s operational environments
- Mining and visualising large structured and unstructured datasets to find new insights to inform operational efficiency and interaction strategies

- Build data infrastructure that scales exponentially year-on-year in terms of both volume and variety as more complex products are introduced into the ecosystem
- Work closely with software engineers and data scientists to enable the delivery of actionable insights in real time; building internal products/tools, utilising open-source software to schedule data jobs, facilitate data flows and enable integration between the core system, the Data Warehouse and other cloud-based systems
- Build world-class, scalable data pipelines and warehouses so that by leveraging the rich data available, we are able to add real value to our customers and drive new product development, both internally and externally
- Support sophisticated predictive data products by maintaining data science production environments (cloud-based, python), ensuring that the outputs from Data Science models are available and integrated into the system and integrate, coordinate and maintain data flows between various sources of data
- Manage and maintain cloud service integrations that perform key data functions by working towards replacing third-party elements of the data pipeline by using open-source tools
- Make decisions around the infrastructure, layout and processes of the data warehouse, including:
    - working with the engineering team on how to best track and record data
    - following up on data inconsistencies to ensure that it is corrected
    - transforming, standardising and collecting data from various sources
    - collecting information from various sources to augment the business tables
    - assisting with the optimisation of SQL queries from the Data and BI teams
    - scheduling and maintaining batch processing jobs

Requirements:

- Completed BSc (Computer Science) degree or similar (Software Engineering, Data Science, Statistics, Operations Research, Applied Mathematics etc with experience in software engineering, computer science or working with big disparate sets of data, statistical modelling, data mining, machine learning or optimisation)
- Other analytical qualifications will also be considered if accompanied by the relevant experience
- *2-5 years experience in the following:*
- Data Engineering
- ETL processes and transformations
- Cloud experience ideally with Google Cloud Platform
- DevOps Stack development experience
- Experience working with large datasets and automated ETLs
- Experience deploying applications to cloud environments
- Experience in using Apache-Airflow or other data pipeline tools
- Familiarity with the structure of data required for reporting and data science projects
- Exposure to Data Processing products common in the eco system - BigQuery, Redshift, Spectrum, S3, Athena, Kafka, Spark, Storm, Flink, Beam, Presto, Hive
- Exposure to Scala or Java in context of data processing
- Experience and proficiency with Python
- Experience in the design and implementation of data flows
- Advanced SQL/PostgreSQL/Redshift knowledge

- Experience with Hadoop and Spark
- Experience in the design and implementation of REST APIs
- Exposure to Google Cloud Platform, Azure, AWS or similar
- Exposure to practical database design
- Experience working closely with machine learning or analytics teams
- Solid coding and programming skills with SQL, and at least one of R, Python, Spark or similar
- Proficient understanding of distributed computing principles utilising Hadoop and MapReduce
- Strong analytical and statistical / machine learning skills
- Ability to formulate problem statements and develop a plan for tackling the problem
- Above average ability to work with, analyse and communicate findings from data (verbal and written)

You can expect:

- Flat hierarchies, a great team and transparent communication
- Collaborative team structure
- Cloud computing – imagine unlimited computing power
- Cutting edge tech stack – Kubernetes, Docker, Rkt, apache beam, pyspark, traefik, BigQuery, BigTable, Cloud Spanner or whatever else is the best tool for the job to get done'
- Onsite training – You want to learn Go? Here's a course. You want to go back to varsity to learn Stats? Done.
- An attractive compensation
- Workplace in the heart of Bryanston
- Open bar
- Pinball, Foosball, mini golf and arcade games in the office
- Quiet rooms for really getting stuff done
- An opportunity to work on the bleeding edge of tech and cloud computing

About the Company:

DotModus has been a Google Partner since 2011 and are now a Google Cloud Premier Partner with a specialization in Data Analytics. Our belief and culture is that your workplace, and the people you work with, should both be amazing. Heck, you spend over a third of your time at work, it better be fun! Work hard, Play hard.

At DotModus, you'll be working with some of the smartest people in South Africa, where you'll get to play and develop with some of the coolest tech and gear we can lay our hands on.

Data is what we do and we're loving it. Our products, projects and customers will expose you to some of the biggest and best datasets around and throw you into one of the biggest growing industries of today. We also build beautiful apps that are as much experiences as they are tools. We build for people, not robots.

Oh, we also drink beer on Friday, did we mention that?

See you soon.

DotModus