Waseem Khalifa

Analytics Data Engineer

Expertise: SQL | Python | dbt | Snowflake | Airflow
Contact: 07706 906209 | waseem.khalifa.engineer@gmail.com
Education: BSc (Hons) Computer Science, 2:1
Location: Coventry, West Midlands, CV2 4GZ
Additional Details: Full UK Driving Licence • Car Owner • British Born Citizen (UK Right to Work)
Notice Period: 3 Months

PERSONAL SUMMARY

A seasoned computer science graduate with over 10 years of experience in the data field, currently serving as an Analytics Engineer at a fast-growing fashion media start-up. My role centers on building robust data pipelines to ingest data from diverse sources into a centralized data warehouse, transforming data using SQL, and leveraging dbt to implement a Test-Driven Development (TDD) approach, ensuring high data quality. This foundation supports the delivery of actionable business insights and the development of scalable reporting infrastructure in Looker. In recognition of my commitment and achievements, I was awarded the 'Star Employee of the Year' award in a previous position, a testament to my dedication and contributions.

To explore my work and expertise in greater detail, feel free to visit my GitHub portfolio:
https://github.com/waseemkhalifa/portfolio

KEY SKILLS & EXPERTISE

- **SQL:** 15+ years of experience writing advanced SQL queries, including window functions, CTEs, UDFs, and complex joins. Skilled in optimizing queries for code readability, performance and cost efficiency.
- **Cloud Platforms:** Hands-on experience with AWS (S3, Redshift), GCP (BigQuery), and Snowflake for data storage and processing.
- **Data Modeling:** Expertise in designing and implementing data models (Kimball, OBT, Data Vault) for data warehouses such as Snowflake, BigQuery, and Redshift.
- **dbt:** Proficient in using dbt to modularize and centralize analytics code, build data marts, and implement TDD practices to ensure data quality and integrity.
- **Scripting [Python & R]:** 7+ years of experience in advanced data manipulation, scripting, and visualisation using Python (pandas, NumPy, scikit-learn) and R (ggplot2, dplyr).
- **Programming [Python]**: With over 7 years of professional experience using Python, have recently expanded expertise to include developing robust applications, such as CLI programs and 2D games. Proficient in advanced programming concepts like OOP, inheritance, polymorphism, functional programming, and data structures and algorithms. Skilled in extracting data from APIs and performing web scraping to efficiently gather and process information.
- **ETL/ELT Tools:** Experience with Airflow, PySpark, Fivetran and Stitch for building and managing end-to-end data pipelines.
- **Version Control:** Daily use of Git for version control, including branching, merging, PR reviews and collaborative workflows.

EMPLOYMENT EXPERIENCE

Analytics Engineer | The Business of Fashion, London (Remote) | Dec 2022 – Present
[Tech stack: SQL, dbt, Snowflake, Looker, Fivetran, Snowplow & git]

Key achievements:

- Designed, developed, and maintained nearly 1,000 automated dbt models using SQL and Jinja, strategically selecting appropriate materialisations (incremental, view, ephemeral, snapshot) to optimize cost efficiency, enhance code readability, and simplify maintenance.
- Implemented a robust Test-Driven Development (TDD) approach within dbt, creating comprehensive tests using SQL, Jinja, and the Great Expectations package to ensure data quality and integrity.
- Addressed Slow-Changing Dimensions (SCD) challenges by developing automated snapshot models in dbt, preserving critical historical data and minimizing data loss.
- Managed end-to-end data transformation pipelines, including:
    - Ingesting new and existing datasets into Snowflake via Fivetran.
    - Transforming raw data into structured models (Source, Staging, Intermediary, Mart, etc.) using dbt, complete with rigorous testing.
    - Developing LookML models to support reporting and analytics in Looker when required.

- Built automated pipeline tests to run post-merge, ensuring that new code changes do not introduce errors or disrupt existing functionality. These tests were designed to detect anomalies, validate data transformations, and confirm that all business logic is preserved, thereby maintaining the integrity of the data ecosystem.
- Developed and deployed freshness tests to monitor data pipelines, ensuring that models were updated within expected timeframes and remained aligned with business requirements. This proactive approach minimized the risk of stale data and unnecessary cost inefficiencies, ensuring that stakeholders always had access to the most current and relevant data.
- Oversaw the GitHub repository for the data engineering team, conducting thorough code reviews to ensure adherence to agreed-upon guidelines and best practices, thereby safeguarding data integrity and minimizing risks.
- Designed and developed a series of sophisticated dbt (data build tool) models incorporating intricate business logic to generate actionable insights and drive strategic decision-making for the organization's subscription service. These models analyzed key performance metrics, including customer journey flow performance on the website, Monthly Recurring Revenue (MRR), and multi-touch attribution, enabling stakeholders to optimize customer acquisition, retention, and revenue growth.
- Transformed raw data into structured, business-ready models that powered comprehensive reports and dashboards, providing visibility into critical subscription metrics and customer behavior. By embedding advanced analytics into the data pipeline, supported data-driven strategies that enhanced the overall performance and profitability of the subscription service.

Senior Web Analyst | Next, Leicester | Feb 2021 – Nov 2022
[Tech stack: Google Analytics, BigQuery, SQL, R, Python, Looker Studio & git]

Key achievements:

- Designed, developed, and deployed highly scalable and automated data pipelines within GCP BigQuery to process and transform terabytes of Next's Google Analytics web data. These pipelines enabled streamlined reporting, advanced analytics, and real-time insights, supporting data-driven decision-making across the organization. Utilized advanced SQL techniques to author complex queries, transforming raw, unstructured data into business-ready tables with embedded logic, ensuring seamless integration with Looker Studio for intuitive reporting and visualization.
- Enhanced data transformation efficiency by leveraging BigQuery User-Defined Functions (UDFs) to optimize pipeline performance, reducing processing times and operational costs. Ensured the scalability and reliability of pipelines to handle large-scale data volumes, maintaining high performance and uptime even during peak data ingestion periods.
- Spearheaded the creation of analytics-driven data models and reports tailored to Next UK/International and their Total Platform service, providing actionable insights that laid the foundation for the Web Analytics team's success. Revamped metric definitions to align with business objectives and introduced robust coding practices, including optimized data scheduling and dataset refresh processes, to improve data accuracy and timeliness.
- Refactored a large part of the existing code base for the web analytics team by implementing clean, modular, and well-documented code using advanced SQL techniques such as Common Table Expressions (CTEs) and UDFs, fostering a culture of knowledge sharing and collaboration. This approach enabled junior analysts to quickly onboard, understand, and contribute to high-value analytics projects, driving efficiency across the team.

Senior Analyst | Spark44, Birmingham | Jan 2020 – Feb 2021
[Tech stack: Google Analytics, BigQuery, SQL, R, Python & Looker Studio]

Key achievements:

- Working on the JLR (Jaguar Land Rover) account, spearheaded a data modeling initiative in BigQuery utilizing terabytes of Google Analytics data, serving as the lead architect for designing and building data marts to centralize key business metrics, ensuring a single source of truth. These automated and pipelined data marts were subsequently adopted by analysts for advanced analysis and business intelligence reporting.
- Developed innovative data marts to calculate session-level engagement scores for users on Jaguar Land Rover (JLR) websites, based on user interactions and behaviors. These engagement scores were instrumental in driving personalization strategies, A/B testing, and performance reporting. This initiative aimed to replace a costly third-party personalization tool, delivering significant cost savings - a critical achievement during the COVID-19 pandemic, which had led to declining sales and heightened financial constraints.
- Took ownership of a high-priority data pipeline project during a contractor's absence, delivering a functional solution ahead of schedule. Designed a Python script to extract site speed metrics from the Jaguar Land Rover website, ensuring accurate data collection. Built a CLI tool to retrieve raw data, followed by transformation, cleansing, and normalization using Pandas. Integrated the processed data into GCP BigQuery, enabling automated reporting dashboards for performance monitoring. This eliminated a

weeks-long backlog, creating a scalable and reliable pipeline, and improved the organization's ability to analyze site speed. Demonstrated strong problem-solving, technical expertise, and timely delivery under pressure.

Insight Manager | UK Flooring Direct, Hinckley | Oct 2016 – Jan 2020
(Previous roles: Digital Marketing Analyst: Oct 2016 – Jul 2017)
[Tech stack: Google Analytics, R, Python, Tableau, Redshift, S3, Stitch, Airflow, Netsuite, Hotjar & Domo]

Key achievements:

- Designed and implemented end-to-end automated data pipelines by ingesting data via Stitch into Amazon S3 and Amazon Redshift, ensuring seamless data flow and storage. Then utilising advanced SQL to develop structured data tables and data marts with embedded complex business logic, enabling actionable insights for business stakeholders.
- Automated Tableau reports leveraging the processed data in Amazon Redshift to support trading strategy development, performance understanding, and data-driven decision-making.
- Temporarily managed and maintained the business's Apache Airflow infrastructure during the data engineer's absence, ensuring uninterrupted pipeline operations and data availability. Demonstrating adaptability and technical proficiency by stepping into a critical role, maintaining system reliability, and supporting business continuity.
- Delivered high-impact insights by analyzing the end-to-end customer journey, from initial website interactions to contact center transactions, through the integration of multiple online and offline data sources. These insights provided the business with a deeper understanding of customer behavior, dispelling previous misconceptions and directly influencing a strategic shift in trading approach.
- Designed and developed multiple Machine Learning models using R and Python, including a classification propensity model to predict lead conversion likelihood, a basket recommendation algorithm leveraging the Apriori technique, and a regression model to forecast a customer's potential order value based on their free sample order history. These models provided actionable insights to optimize marketing strategies and enhance customer engagement.

Online Marketing Executive | Soak.com, Nuneaton | Jul 2013 – Oct 2016
(Previous roles: Marketing Assistant: Sep 2013 - Apr 2016, Internship in Marketing: Jul 2013 - Sep 2013)
[Tech stack: Microsoft Excel]

PERSONAL PROJECTS

End-to-End Airflow & dbt project with custom ecommerce data
Link: https://github.com/waseemkhalifa/portfolio/tree/main/Data%20Engineering/ecommerce_pipeline
Tech used: Airflow, dbt, Python, SQL, AWS S3, AWS RedShift
Summary: Designed and implemented an automated data pipeline using Apache Airflow, encompassing the following key components:

- Developed a Python program to extract and store synthetic e-commerce data in nested JSON format in an AWS S3 bucket.
- Engineered Python scripts to extract, transform, and load data from S3 into AWS Redshift.
- Utilized dbt to create structured data models (dimension, fact, and mart tables) in Redshift, optimizing data for business insights and reporting.

End-to-End Airflow & PySpark project with custom ecommerce data
Link:https://github.com/waseemkhalifa/portfolio/tree/main/Data%20Engineering/ecommerce_pipeline_pyspark
Tech used: Airflow, PySpark, Python, AWS S3, AWS RedShift
Summary: Much like the project above, built a automated Airflow ETL pipeline but utilised PySpark for data transformation to create dimension and fact tables

IMDB movie analysis using APIs and Web Scraping
Link: https://github.com/waseemkhalifa/portfolio/tree/main/Data%20Analysis/My%20IMDb%20Ratings
Tech used: Python, R, API, Web Scraping, Tidyverse, ggplot2, pandas, dataclasses, requests, BeautifulSoup
Summary: Performed an in-depth analysis of over 1,000 films rated on the IMDB movie database by leveraging APIs and web scraping techniques with Python, complemented by advanced data visualizations created using the R programming language.

EDUCATION: BSc Computer Science | 2.1 | Coventry University | 2009 - 2012

- OOP (Object Oriented Programming) in Java, SQL, Database Design etc

REFERENCES: References available upon request.