KARANPREET KAUR

Data Engineer

Toronto, ON   +1 (778) 773-5498   karanpreet088@gmail.com

https://www.linkedin.com/in/karanpreet-kaur-088/

Technical Skills

Python | Spark/Pyspark | Azure Data Factory | Databricks | Data Warehouse / Synapse Analytics | Data Pipelines | Databases & SQL | MongoDB | Data Modeling | Azure Event Hubs | Hadoop | Big Data | Predictive Modeling | Machine Learning

Non-Technical Skills

Communication | Leadership | Presentations | Mentoring

Work Experience

Avanade (Established by Microsoft & Accenture**)** — **July 2022 - Present**

Data Engineer – Intelligent Data Platforms (Toronto, ON)

- Developed **end to end data extraction pipeline** to ingest data from different database sources, it's respective config tables, handles schema evolution and outputs parquet files ready to ingest in delta lake
- Led POC to data profile 1K+ high priority synapse tables and implemented partitioning and indexing strategies **and reduced query times up to 60%**
- Designed and implemented loading some entities hit by synapse storage limit as external tables which made these tables **available for end user access**
- Developed custom **orchestration** job onboarding ETL pipeline which replicates the on-prem oozie workflows
- Build Spark Python script to compare delta ingested and on-prem partitions for all active entities which helps team to regularly evaluate effected partitions with data ingestion issues
- **Point of contact** to communicate with source teams to resolve any data discrepancies and any source data changes
- Contributing to company's internal business development - supported running the Analytics Technical Recruitment Process optimization thread, member of Avanade Reads

RocketBrew — **May - June 2022**

Data Scientist + Data Engineer (Vancouver, BC)

- Understand business requirement to classify product images into categories to help their customers monitor their online competitors, developed and presented **proposal** for data science techniques to achieve it
- Performed **EDA** on the product data, removed duplicates and their distribution across major categories
- Showcase data findings, data science model techniques and model results in weekly meeting with CTO
- Developed an unsupervised machine learning model to **classify ~200K+ products** from different stores across various ecommerce platforms into their subcategories. The methods CLIP and multiclass classification were ensembled/combined to achieve **higher precision** for each category of products.

Deloitte India LLP — **Aug 2018 – July 2021**

Data Analyst + Data Engineer (Gurugram, India)

- Identified and removed redundant activities in the execution workflow in Azure Data Factory, which leads to a **reduction of 1hr in daily execution and 45 mins in the monthly process** and also decreased consumption of cloud resources

- **Reduced** storage and processing time in SQLDW by analyzing the duplication of records between spark and SQL layer, and hence reduced the **row count by 86%**
- End to end implementation and **automation** of ETL process **expedited deliverables by 1-2 days** every month and made team independent of any external and manual dependencies for deliverables of Power BI dashboards
- Led **design, development** and **validation** of external data source dashboards, represented as single point of contact from team for any process related queries
- Replicated **complex SQL queries** implemented in SQLDW in Spark (Azure Databricks) which saved 5 hours of execution time and cut down 650GB storage in the data warehouse
- Implemented entity extractor model for a POC on Chatbot engine bottom-up using RASA stack and integrated with actions module to cater financial queries for an Investment Bank, in turn, enabling **cost reduction of 4-5 FTE's**
- Assisted in making decks for client proposals on Finance Chatbot propositions
- Self-initiated and developed a POC to transform structured data into the natural language using ARRIA Studio (ATL language), which automates the manual efforts and savings of **1 FTE** spent on writing commentaries for Tax and Revenue reports generated every month

Portfolio Project

Online Taxi Service ETL Project:
Developed an ETL pipeline process that generates data for an online taxi service database and weblogs. It handles erroneous data and tracks ETL metadata using logging, including job start time, job finish time, and status. By performing data wrangling, it transforms it into a readable data format for reporting and finally fills the initial load of the target datastore. (GitHub)

Education

- Masters of Data Science, 2021-22 (*The University of British Columbia, Vancouver, BC*)
- Bachelors of Engineering in Computer Science, 2014-18 (*Thapar Institute of Engineering & Technology, India)*

Certifications

- Azure Data Engineer Associate (DP-203)
- Databricks Certified Data Engineer Associate

Honors & Awards

- 3x Hero Award for Exemplary Performance, Avanade 2022-23
- Live the Dot, Deloitte 2020 - Awarded for Consistent performance/outstanding contribution
- Move the Dot – Team, Deloitte 2019 - Awarded for exemplary performance/ significant contributions