ANSHU ARYA

Senior Data Engineer

B.tech (InformationTechnology)
tech.anshu09@gmail.com
+91-8553019693
LinkedIn/anshu-arya

Following the passion to solve real world problems through data and technology

EXPERIENCE (6 Years)

Publicis Sapient, Bangalore — *S*enior *Data Engineer*

April 2022 – PRESENT

- Built **ELT** pipelines using **Python** and **DBT**
- Managing and scheduling the processing of data in the sandbox.
- Worked with large dataset included structured and unstructured data
- Created **codebuild** pipelines and orchestrated on **Airflow**
- Designed and implemented the code using Python to parse the raw files , processed using DBT and loaded into **Snowflake**
- Experience with **Agile Development** and the **day-to-day scrum**.
- Worked on client facing environment

Cerner Healthcare, Bangalore — *S*enior *Software Engineer*

JAN 2019 – March 2022

- Collaborated with various teams & management for migrating data from relational database (**OracleDB**) to Cloud warehouse (**Snowflake**).
- Used **SQL** to analyze table data in Oracle DB.
- Worked with **Cloudera 5.16** and its different components (**Hadoop, Hive, Hue, Oozie**)
- Built **ETL** pipelines using **Python** and **Apache Spark**
- Triggered ETL Jobs on on-Prem Hadoop clusters and **AWS EMR**.
- Designed and implemented the code using Java to fetch the oracle table schema and convert it to the **Snowflake warehouse schema**.
- Worked on on-call production issues - Resolve snowflake schema issues, workaround for **defects** within SLA duration.

Quess Corp, Bangalore — *Associate*

Mar 2017- Sep 2018

- Collaborated with the Data Science team and parsed server application-generated logs using **Java** to gather useful information (Flight event info).
- Created the Java class that fetched, parsed, and formatted the Flight server logs, and stored Flight event info into the **SQL** database.
- Built an **API** that follows **RESTful** standards to Perform **CRUD operations** on stored data sets in a database.
- Migrated existing code base to Java from Ruby.

SKILLS

PROGRAMMING LANGUAGES

- **Python, SQL**

AWS & SCHEDULERS

- **AWS (s3 ,Ec2, EMR, SNS, Cloudformation, Cloudwatch,lambda)**
- **Azure**
- **Airflow,oozie**

DATABASES

- **OracleSQL, SqLite.**

BIG DATA TECHNOLOGIES

- **DBT**
- **Hadoop (CDH)**
- **Hive**
- **Snowflake**
- **Py Spark.**

CI/CD

- **Docker**

FILE FORMATS

- **Avro, Json, csv.**

CONTROL SYSTEMS AND DOCUMENTATIONS

- **Git, Jira, Bitbucket**

OPERATING SYSTEMS

- **macOS, Ubuntu, Window**

LANGUAGES

- **English, Hindi**

,

PROJECTS

FI-Daily-Process-ELT — *Franklin Templeton (April 2022 – PRESENT)*

DESCRIPTION :

- FI-ops team sends FT data to port which is loaded in port using Python splitter, parser and loader service.
- This data is then processed using DBT
- The Data Sandboxes is a scalable and developmental platform to explore an organization rich information set.
- The data sandboxed are created by specifying database objects(Tables,views,schema,user groups etc)
- The code is complied using codebuild artifacts are deployed using airflow and data is loaded into FT snowflake databases

Analytics-millennium-ETL— *Cerner Healthcare(JAN 2019 - March 2022)*

DESCRIPTION :

- The Crawler service runs as a set of parallel threads performing ordered tasks that facilitate detection, discovery, extraction, and uploads of the changed Oracle table's data, monitored by a tracking system and orchestrated by a fair-share scheduler.
- The crawler creates a small number of JDBC connections to the database and uploads extracted content to the Wolfee data collector service, which stores the data in a distributed file system in an Avro file format (Wolfee events).

Analytics-millennium-ETL— *Cerner Healthcare(JAN 2019 - March 2022)*

DESCRIPTION :

- Analytics-millennium-etl ETL for processing Oracle table data into target warehouses (Snowflake).
- The purpose is to replicate clients' Oracle table's data in their corresponding Snowflake warehouse schemas.
- Analytics-millennium-etl takes in a client's Wolfe Source ID and logical domain ID, and it retrieves WolfeEvents from one of the long-term storage solutions based on a provided date range and the client's Wolfe Source ID.
- Once the qualifying data are retrieved from storage, the analytics-millennium-etl then deserializes the WolfeEvents and further qualifies the data on the client's logical domain ID.
- The data is then deduplicated based on the target table and key, such that the latest row for each table and key is retained and the rest is discarded.

(IFDT) In-Flight-Data-Transfer tool — *Delta Airlines*

MAR 2017 – SEP 2018

DESCRIPTION :

- IFDT (In-Flight & Data Transfer) Application, which will enable the transfer of different System Generated Logs in Flight Server in the Cloud deployed Ground Analytics Server.
- The IFDT Application monitors different events occurring in Flight (FO, FC, PO), controls and transfers the data generated between every pair of events to the ground-based analytics system through a secured connection.
- Also, parsing a few data into a human-readable format (Excel, CSV).

EDUCATION

University College of Engineering and Technology (Vinoba bhave University), Hazaribagh — *B.Tech (Information Technology)*

2012- 2016