Prakhar Vishnoi

Gurugram prakhar2170@gmail.com +918837702899 prakhar2170

About

An Engineer and initiator with analytical mindset that aims towards business solutioning. A developer with strong technical and conceptual foundations that enables efficient business implementations. Looking for a niche tech oriented role in Data Engineering profile involving scalable designs and architectural solutions.

Summary

As Data Engineer

- Experience as a Tech Lead (Data Engineer) for **Gilead's Data Engineering in-house platform** built on AWS Cloud**,** called **CCF (Common Components Framework)**
- Worked with CCF platform architects in solving design problems like AWS **CloudWatch** and **EMR API Throttling**, **AWS MWAA Airflow** issue and **AWS RDS** contention performance issues.
- Experience in creating **PySpark DataWarehousing Facts and Dimension scripts** for execution on **AWS EMR**.
- Experience in Configuring **AWS Appflow** to ingest data from **Salesforce API**.
- Experience in using **Boto3** AWS API for **AWS S3, Appflow**, etc. services.
- Experience in ingesting data into CCF from various external **JDBC** sources like **MySQL, Oracle** etc.
- Experience in handling complex data types in SQL and pyspark using **json functions (from_json, schema_of_json) and other functions (collect_set, flatten, array_distinct)**
- • Experience in using Databricks UI to create Workflows, Jobs and Repos.

As ETL Developer

- An ETL developer with immersive experience in **Informatica, SQL Server and PL/SQL.**
- Used **Informatica Client tools** namely **Power Center Designer, Workflow Manager, Repository Manager and Workflow Monitor**.
- Created Informatica **mappings, mapplets and transformations** (Stored procedure, Union, Filter, Router, Update Strategy, SQL, Joiner, Aggregator, Expression, etc.).
- Created **sessions and scheduled workflows** in Workflow Manager, Informatica PowerCenter.
- Comfortable with both MS Windows and Linux environments. Experience in **Windows Batch Scripting**.
- Extensively used DDL, DML commands, complex joins, triggers, functions, stored procedures, rank function, synonyms, etc. in **MS SQL Server (T-SQL).**
- Created re-usable **Oracle PL/SQL Procedures and Functions** to perform inter-database Data Migration.
- Automated Server Audit into a tool in PLSQL using cursors (nested looping and recursion), views, indexes (clustered and non-clustered) and temporary tables.
- Executed Server Audit Tool across multiple servers using **Windows Batch Scripts**.

Experience

National Australia Bank, Senior Data Engineer, Gurugram (May 2023 - Ongoing)

The ADA Platform follows databricks recommended design and had data loaded into 3 layers –Bronze, Silver and Gold. Loaded data into ADA bronze from multiple files based json, csv and high-volume resources HVR sources. Sources were migrated from old AWS based platform NDH to Databricks based ADA

- Requirement gathering, feasibility analysis and design data pipeline for new ingestion.
- Created Databricks Workflows consisting of DLT jobs to load data into target tables from S3 linux mounted locations.
- Created incremental batch and realtime pipelines using cloud_files and stream functions.

- Loaded data from kafka generated json datasets, selecting key and value columns and parsing value columns using, from_json, schema_of_json functions.
- Worked on complex datatypes using collect_set, flatten, array_distinct functions.
- Created Pyspark and SQL based notebooks toExtract load and transform and bulk migrate history tables to ADA Platform.
- Create estimation and get signoff from Design forum for new requirements.
- Guiding and leading team to develop the platform to create data pipeline for file based and HVR using Databricks platform.
- Created Python and SQL UDFs on databricks while performing data profiling on Databricks Notebooks.

ZS Associates, Senior Data Engineer, Gurugram (May 2021 - May 2023)

Gilead AWS Cloud CCF (Common Component Framework) EDP (Enterprise Data Platform)

ZS Built and supported Gilead's primary Cloud Enterprise Data Framework called CCF. CCF, an enterprise datawarehousing Platform, had the capabilities to ingest the data from external sources and maintain a persistent DataWarehouse on AWS S3. CCF uses PySpark and Hive for compute, And is adopted by multiple business LOBs consisting of more than 7000 developer/user base.

- Lead a team of 10 members across 3 time-zones IST, EST and PST for **entire** CCF Production Support with absolute trust from the client. Received positive feedback from the client in terms of Delivery, Team/People Management, and fixed some of the most difficult technical challenges.
- Configured AWS MWAA, a relatively new managed AWS service for Apache Airflow, for usage across the CCF Platform. Eliminated NEGSIGNAL, SIGTERM and performance issues in Airflow.
- Eliminated AWS API (CloudWatch and EMR) Throttling issues from CCF platform using Exponential Backoff Retry technique.
- Fixed AWS RDS contention issues, on CCF control tables, by determining appropriate index sequence and cluster size for performance optimization.
- As a part of CCF Platform Team, Worked with other Data Engineers from LOBs to help them fine-tune their business logic as per CCF design and limitations. Lead a team to fix the minor bugs in CCF and played a key role in the evolution of CCF.
- Performed a POC to integrate Redshift Data warehouse tables from downstreams to AWS GlueCatalog, to viewed in Athena as a part of Datalake.
- POC: Designed and implemented a Jenkins-based CI/CD pipeline for automated testing and data validation in AWS, utilizing Pylint for static code analysis and Great Expectations for data quality checks in data engineering workflows.

Roche GNE Apache HUDI POC

POC, Roche GNE aspire to move aspirations to move towards Apache HUDI instead of native PySpark. POC involves SCD1/2 implementation on HUDI.
Apache HUDI is a Data Lake solution, A Databricks Delta open-source alternative, has support for real-time BigData streaming along with support for performing CRUD operations directly on distributed storages (e.g. AWS S3) and persistent filesystem storage on disks. Apache HUDI is adopted in market by Disney+Hotstar and Tencent for their BigData streaming systems that have PetaBytes of data.

Gilead MDH (Marketing Data Hub) on EDP

Gilead's in-house Marketing Data Hub for Data Science and Analytics. A Data Engineering Engagement which required custom changes over CCF and building ingestion and Data Warehousing pipelines in Gilead's DataLake for marketing data ingestion.

- Goto person for CCF understanding in MDH Business team. Lead all technical issues and tasks.
- Fixed a Design Implementation of datasource Ingestion in CCF Platform, while working in a Business Team.
- Optimized AWS AppFlow ingestion Utility, to prevent runtime errors while ingesting large volumes of data from salesforce, in production.
- Fixed AWS RDS contention issues, on CCF control tables, by determining appropriate index sequence and cluster size for performance optimization.

- Created a Unique and specific Technical Design for aggregate and non-aggregate FACTS that override the default behavior of DW creation in CCF. Enabled this business requirement design which was unsupported by CCF Platform. Lead a team of 8 members and onboarded the team to CCF platform.

Gilead Medical Affairs on EDP
Gilead's in-house Marketing Data Hub for Data Science and Analytics. A Data Engineering Engagement which required custom changes over CCF and building ingestion and Data Warehousing pipelines in Gilead's DataLake for marketing data ingestion.

- Ingest data from 5 different datasources (Salesforce API using AWS AppFlow, Oracle and Teradata JDBC Sources, FTP server and Http Web servers) into Gilead's CCF Datalake.
- Developed an Http based Upstream datasource ingestion utility for CCF, using Python and Urllib module.
- Developed an AWS Appflow based Upstream datasource ingestion utility for CCF, using Python and Boto3 Appflow API.

ITC Infotech, ETL Developer, Bangalore (July 2018 - Nov 2020)

Comprehensive Healthcare Assessment (CHA)
Sutter Health's in-house replacement tool to Voyager RA service provided by vendor Dynamic Healthcare Systems.
Voyager RA: Online automated workflow tool that determines eligible claims for coverage under the insurance coverage contract as per guidelines set by CMS.
Dynamic Healthcare Systems: Business solutions provider that implements processes used by other organizations as online services for Medicare Advantage and Commercial Exchange insurance plans as per CMS guidelines.
CMS: Centers for Medicare and Medicaid Services is a US Federal Agency that works in partnership with state governments to administer Medicare and Medicaid services.

- Single-handedly managed the development at offshore with ownership in both design and development.
- Designed **ETL pipeline architecture** and **implemented in Informatica PowerCenter Mapping Designer**.
- Made efficient delivery through **change suggestions and recommendations** in **Architectural Design Documents** as per business requirements from offshore team.
- Created **source and target datastores (tables**) in development environment (SSMS/SQL Server).
- Created **proper keys and indexes** in relational datastores, for **performance optimization of ETL pipelines**.
- Coded the logic (**Created Mappings**) for claim processing in Informatica Mapping Designer specified in design document. **Used various Informatica Transformations** (Stored procedure, Union, Filter, Router, Update Strategy, SQL, Joiner, Aggregator, Expression, etc.) to transform the data.
- **Created sessions** for mappings, **configured session properties**, **created workflows and scheduled Workflows** in **Workflow Manager** for frequent runs.

Enterprise Data Management and Audit Tool
Automated the Server Audit process (manual earlier) single-handedly creating a reporting tool that defined the Data Landscape of all Database Servers, one at a time, all in parallel.
Enterprise Data Management is an ever-running activity that targets controlled accessibility over entire organization's data from an enterprise standpoint. This project defined the entire Enterprise **Data Landscape** enabling data cross-availability across various endpoints for applications and analytical uses. The Data Landscape definition exposed the static and unused data in operational/production datastores which was decommissioned for storage cost cutting. This was achieved through **Data Rationalization** (marking the data as qualified or non-qualified for its existence on production Data Stores) which marked 70% enterprise's data as non-qualified thus enabled **Data Decommissioning**. **CCPA** effective from 1st Jan, 2020 required enhanced **Data Governance** for consumer data protection, which required **Data Classification** (identification of PHI/PII and other consumer specific sensitive information stored in Data Stores) to mark data with consumer permissions for collection, access, storage, disclosure and sale of data.

- Had frequent client interactions, discussions with manager to suggest to client - cost cutting and governance enabling methodologies, followed by **implementation of a constantly evolving Audit Tool** in SQL Server.
- The Tool was coded in SQL Server. Tool had two major parts:
- Data Rationalization for Data Decommissioning.

- Data Classification and Risk Attribution for Data Governance.
- **Created Windows Batch Script** for parallel **execution of Audit Tool across multiple Database Servers** specified in a Configuration File.
- **Documented the Functional specifications** for the above.

Education

B.Tech. Lovely Professional University, Computer Science and Engineering — August 2014 – May 2018

- GPA: 9.12/10
- **Coursework:** Object Classification using Deep Convolution Neural Network, AlexNet via Transfer Learning and hyper-parameter tuning.

Skills and Technologies

Languages: Python, Core Java (Intermediate), C++, C (Strong Basics)

Technologies:

- PySpark
- AWS (S3, EMR, EC2, Glue, MWAA, Lambda, CloudWatch, Athena, IAM, RDS, Appflow)
- Hive
- Putty
- Unix Shell
- Microsoft SQL Server (T-SQL) 2008, 12, 14
- Oracle PL/SQL
- Apache HUDI, PostgreSQL, Teradata, WinSCP, PuTTy
- SQL (TSQL - DDL and DML) and Informatica PowerCenter 9.6
- Windows Batch Scripting, PLSQL (TSQL) and SQL (TSQL)